ABSTRACT

Title of Dissertation: **COMPARATIVE GENOMICS OF CHEMOSENSORY GENE FAMILIES AMONG MEMBERS OF THE HELIOTHINAE**

Rong Guo, Doctor of Philosophy, 2023

Dissertation directed by: Dr. Megan Fritz,
Department of Entomology,
University of Maryland, College Park

Insect chemosensory systems play crucial roles in the perception of chemical signals that regulate sexual behaviors and odors that mediate insect-host plant interactions. These processes, mate-finding and acceptance, as well as host plant identification and use, strongly contribute to diversification and speciation among plant-feeding insects, including the Lepidoptera. *Chloridea virescens* and *Chloridea subflexa* are an ideal pair species to study the evolution of insect chemosensory systems because they are closely related but show pheromone-based sexual isolation and divergent host plant preferences. My dissertation focuses on the development of genomic tools that enable investigation into genetic mechanisms of host plant and mate recognition, and applies these tools to examine inter- and intraspecific diversity of chemosensory genes among members of the Heliothinae.

In chapter 2, I produced a novel Illumina short read *C. subflexa* genome assembly and an improved, highly contiguous *C. virescens* genome assembly. Due to quality limitations common to short read assemblies, I used our Heliothine genomes to examine the feasibility of reference-assisted assembly, an approach that leverages existing high quality genomic resources for genome improvement in closely related taxa. My work demonstrated that reference-assisted assembly has the potential to enhance contiguity and completeness of existing insect genomic resources with minimal additional laboratory costs. Both the potential and pitfalls of reference-assisted assembly are discussed in light of my results.

In chapter 3, I manually curated two chemosensory gene families, the odorant receptors (ORs) and odorant binding proteins (OBPs), in *C. virescens.* In total, I identified 80 ORs, 1 Orco and 49 OBPs. Three types of OBPs were identified according to the number and positions of conserved cysteine residues: 34 classic OBPs, 8 Minus-C OBPs, and 7 Plus-C OBPs. In addition, I used phylogenetic analyses to study evolutionary divergence of OR and OBP gene families among Heliothine moths, which revealed both gene duplications and losses.

In chapter 4, I studied the strength and nature of selection on the ORs of field-collected *C. virescens* and *C. subflexa*, with focus on the pheromone receptor genes. I characterized the host plant use of these species in central Maryland by comparing the larval densities and infestation rates in 2020 and 2021. Sequencing followed by analysis of selection on field-collected samples indicated that the pheromone receptor, OR6, was under very strong purifying selection in both *C. virescens* and *C. subflexa*. AMOVA tests suggested that in *C. virescens*, host plant-associated population differentiation existed in genes OR6, OR55, OR66 and OR78. Further analyses of genetic divergence analysis focused on OR6 showed that the most highly divergent sites were all in introns. The new genomic tools and analyses of chemosensory gene

families described here will serve as a platform for future investigations into the genetic mechanisms underlying host plant specialization and sexual communication among lepidopteran insects.

COMPARATIVE GENOMICS OF CHEMOSENSORY GENE FAMILIES AMONG MEMBERS OF THE HELIOTHINAE

by

Rong Guo

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Dr. Megan Fritz, Chair
Dr. Anahí Espíndola
Dr. David Hawthorne
Dr. Astrid Groot
Dr. Stephen Mount, Dean's Representative

Preface

This dissertation is composed in part of previously published work, included here as Chapter 2 with the recommendation of the dissertation director and dissertation committee members. The citation for this publication is as follows:

Guo, Rong, Alexie Papanicolaou, and Megan L. Fritz. "Validation of reference-assisted assembly using existing and novel Heliothine genomes." *Genomics* 114.5 (2022): 110441.

As directed in the graduate catalog, I state that I was responsible for the inception of the manuscript and the majority of manuscript preparation. This publication was reformatted to meet university guidelines and supplementary material was placed in appendix, but the publication has otherwise been reproduced exactly. The publication is cited throughout the dissertation, where appropriate.

Dedication

I dedicate the thesis to my dear little boy, Finn Jin, who was born in the fourth year of my Ph.D. study.

This is mom's special spiritual wealth for him.

Hope he can find his confidence and courage to overcome difficulties from my story.

Acknowledgements

Seven years' Ph.D. study was a long journey, and it is my pleasure to acknowledge people who were instrumental for completion of my Ph.D. research. Thank you all for helping me see the adventure through the end.

First of all, I would like to express my sincere gratitude to my advisor, Dr. Megan Fritz, for her continuous support, tireless guidance and patience. It has been a great fortune for me to be a member of Fritz lab. I learned a lot from her, not only the enthusiasm and rigor for science, but also the positive attitude to difficulties, even English - sometimes she revised my manuscript word by word. Thanks for her kind encouragement and valuable advice in my Ph.D. study, especially after I failed my first qualifying exam and during my maternity leave.

I would also like to thank the rest of my thesis committee: Dr. Anahí Espíndola, Dr. David Hawthorne, Dr. Astrid Groot and Dr. Stephen Mount, for their insightful comments and valuable advice on my research, and also for their guidance on becoming a young, independent scientist. During my Ph.D. study, I have had the honor to work with many other worldwide excellent PIs, including Dr. David Heckel, Dr. Fred Gould, Dr. Alexie Papanicolaou, etc. I would like to thank them all for their brilliant ideas, instant feedback, patience and collaboration.

I would like to acknowledge all the members of Fritz lab, as well as the former lab members Anna Noreuil and Dr. Schyler Nunziata, for their help in bench work and field collection. It has been my pleasure to work with these passionate, knowledgeable, fascinating and friendly people for 6 years.

I would also like to give special thanks to my parents for their unconditional love and support. Since Finns' birth, my parents have been especially supportive, staying with us for 8 months as we fumbled our way through the early days of parenthood. They came to the United States from China during the pandemic. I have learned so much from them and I appreciate all of the sacrifices that they have made for me.

The last two people I want to thank are my husband Hang Jin and my dear son Finn. Hang is an amazing husband and father. I am grateful to my husband not only because he has given up so much to make my career a priority in our lives, but also because he has seen me through the ups and downs of the entire Ph.D. process. Finn is almost three years old now, and I need to thank my little boy for being such a bundle of joy and laughter.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AL	Antennal Lobe
AMOVA	Analysis of Molecular Variance
CMREC	Central Maryland Research and Education Center
CSP	Chemosensory Protein
B. mori	*Bombyx mori*
Bt	*Bacillus thuringiensis*
CDS	Coding Sequence
CRISR	Clustered Regularly Interspaced Short Palindromic Repeats
*C. virescens/*Hvir	*Chloridea virescens*
*C. subflexa/*Hsub	*Chloridea subflexa*
DdRAD-seq	Double Digest Restriction-site Associated DNA Sequencing
D. melanogaster	*Drosophila melanogaster*
EF-1α	Elongation Factor -1α
EL	Extracellular Loop
GOBP	General Odorant Binding Protein
GPCR	G-Protein-Coupled Receptor
GR	Gustatory Receptor
*H. zea/*Hzea	*Helicoverpa zea*
*H. armigera/*Harm	*Helicoverpa armigera*
*H. assulta/*Hass	*Helicoverpa assulta*
IL	Intracellular Loop

INSDC	International Nucleotide Sequence Database Collaboration
IR	Ionotropic Receptor
JTT	Jones-Taylor-Thornton amino acid substitution model
LD	Linkage Disequilibrium
MP	Mate Pair
MUM	Maximal Unique Match
NCBI	National Center for Biotechnology Information
OR	Odorant Receptor
OBP	Odorant Binding Protein
OSN	Olfactory Sensory Neuron
PBP	Pheromone Binding Protein
PE	Paired End
PG	Pheromone Gland
PR	Pheromone Receptor
Prob	Probability
QTL	Quantitative Trait Locus
RNA-seq	RNA-sequencing or transcriptome sequencing
RT-PCR	Reverse-Transcription Polymerase Chain Reaction
TMD	Transmembrane Domain
VCF	Variant Call Formatted
WGS	Whole Genome Sequencing
WMREC	Western Maryland Research and Education Center

Chapter 1: Introduction

The insect chemosensory system plays a critical role in the perception of chemical signals that regulate sexual behavior and odors that mediate host plant interactions. The processes of mate-finding and acceptance, as well as host plant identification and acceptance are thought to have contributed widely to the number and diversity of plant-feeding Lepidoptera that we see today. Nearly all species (~200,000 species) of Lepidoptera are phytophagous in their larval stages (Chapman, 2009). Host plant adaptation can potentially lead to ecological specialization of populations and subsequent speciation, and the process is thought to be a main reason for the astonishing diversity of phytophagous insects (Jousselin and Elias, 2019). Mate finding and acceptance involves long-distance signaling that is mediated by female sex pheromones in most moth species (Smadja and Butlin, 2009). The species-specific sex pheromones coordinate mate and species recognition and create prezygotic reproductive barriers, leading to speciation (Cardé and Haynes, 2004). These two processes, mate acceptance and host plant recognition, both contribute to the process of speciation and can be linked by insect chemosensory systems.

1.1 The insect chemosensory system

The antenna is the primary organ used by insects for odorant detection, although other body parts (e.g. maxillary palps, legs, ovipositor) can be involved in chemosensation. Moth antennal morphology has been optimized to detect odorants:

olfactory sensory neurons (OSNs) are located within the multiporous cuticular sensilla on the antenna. OSN dendrites extend into the sensilla and the axons project into the antennal lobe (AL). Most olfactory sensilla are arranged in an array on the antennal branches, enabling them to efficiently catch the odorant molecules passing over the antenna (Koehl, 2005). Male antennae have large numbers of olfactory sensilla, and they can be generally categorized into four morphological types: long trichodea, medium trichodea, basiconica and coeloconica (Sakurai et al, 2014). Different sensillum types house different OSNs that are tuned to different odorants, including pheromone components, plant-derived odorants, etc.

In the process of odorant detection, the odorant molecules first pass through the tiny pores of the olfactory sensilla, and some bind to the soluble olfactory proteins (i.e. odorant binding proteins) in the lymph fluid. Hydrophobic molecules (e.g. pheromone components) become water-soluble through binding, and then the molecules are carried to OSN dendrites, and finally bind with the receptors on the membrane of OSN dendrites (Sakurai et al, 2014).

Gene families involved in insect chemosensation include but are not limited to: (1) soluble olfactory proteins that transfer chemicals via the lymph to corresponding chemosensory receptors, such as odorant binding proteins (OBPs) and chemosensory proteins (CSPs); (2) chemosensory membrane proteins, odorant receptors (ORs), gustatory receptors (GRs), ionotropic receptors (IRs). All three of these latter gene families encode ionotropic ligand-gated ion channels. Generally, OR and OBP gene families are involved in the perception of moth pheromones and host

plant volatiles, and their functional role in mate-finding and host plant detection is well understood (Groot et al, 2016; Chen et al, 2020; Zhou et al, 2015).

Previous studies of OR gene function supported the "one-receptor-to-one-neuron" organization in insect chemosensory systems (Vosshall et al, 1999), which means each OSN expresses a single chemosensory receptor that defines its ligand selectivity. Consistent with this organization, the number of the chemosensory receptors should roughly match the number of olfactory glomeruli, as well as the detected odorants. However, recent studies of the mosquito, *Aedes aegypti*, document a different organizational principle, with several OSNs co-expressing multiple chemosensory receptor genes. This non-canonical model explained the broad ligand-sensitivity of mosquito olfactory neurons (Herre et al, 2022). Similar evidence was also found in Lepidoptera, where some pheromone receptors (PRs) can respond to more than one pheromone volatiles (Wang et al, 2011; Wang et al, 2018; Cao et al, 2021).

1.2 Evolution of the insect chemosensory system

The evolution of insect chemosensory genes is characterized by an extremely rapid process of birth-and-death evolution (Roelofs and Rooney, 2003). The number of chemosensory genes is generally associated with the range of odorants a species needs to detect. For example, within the genus *Helicoverpa*, species feeding on broader host plant ranges have more chemosensory genes than species with narrower host plant ranges (Pearce et al, 2017). These inter-specific differences in

chemosensory gene numbers can be explained by lineage-specific expansions and contractions of gene families over evolutionary time (Sánchez-Gracia et al, 2009).

Chemosensory gene families originate from gene duplication and then diverge and evolve independently. Genomic evidence (e.g. the distribution of chemosensory genes, phylogenetic analysis) demonstrates that unequal crossing over is the main mechanism to generate tandem gene duplications of the chemosensory genes (Sánchez-Gracia et al, 2009). The paralogous chemosensory genes may undergo neofunctionalization (i.e., it may acquire a new function), loss of function or complete loss from the genome by deletion. Shifts in chemosensory gene function or odorant perception may occur through point or segment mutations in the coding regions or the regulatory elements (Graur et al, 2016). The process of neofunctionalization is driven by positive selection acting upon some chemosensory genes (Forêt and Maleszka, 2006; Guo and Kim, 2007; Gong et al, 2009). Meanwhile, the duplicated genes may be silenced and become pseudogenes due to deleterious mutations, which is indeed the most likely fate of a redundant duplicate gene (Graur et al, 2016).

1.3 The role of insect genomes in the study of chemosensory gene evolution

Insect chemosensory gene families are large and because they arise by birth and death processes, some genes may show significant sequence similarity. A well-assembled genome provides sequence data for all chemosensory genes, as well as positional information, which helps to distinguish genes with significant sequence similarity from one another. Comparisons between insect genomes can be used to

understand how gene families, such as those involved in chemosensation, have expanded and diversified or contracted across insect lineages.

Comparative genomics is the direct comparison of whole or large parts of the genome of one organism to that of another to better understand the basic biological similarities and differences, as well as the evolutionary relationships between them (Sivashankari & Shanmughavel, 2007). Advances in sequencing technologies and assembly strategies have dramatically improved the quality of recent genome assemblies, and comparative genomics has become more feasible with the availability of high-quality completed reference genomes. A commonly used approach to the study of chemosensory gene family divergence across taxa is comparative transcriptome analysis: sequencing, assembly and comparison of complementary DNA sequences derived from messenger RNAs expressed in the chemosensory structures of insects (RNA-seq). Yet RNA-seq data are often only collected at a specific developmental stage in a specific tissue, and genes that are not expressed at the time/in the tissue would be missing in the transcriptome analysis. Genomic data provides comprehensive information about the entire genome, both the coding regions and non-coding regions, which include introns and regulatory sequences. Genomic data also contains more information about gene structure and distribution, so comparative genomics has been used to find chromosomal rearrangements and the mechanisms underlying the rearrangement events. As an example of the power of comparative genomics, classic genetics revealed only nine inversions that distinguish humans and chimpanzees at the species level, while comparison of their genomes identified approximately 1,500 (Feuk et al, 2005).

Comparative genomics approaches have also been applied to study key gene families in closely related insect species. One example is the recent study of two major lepidopteran pests, *Helicoverpa armigera* and *Helicoverpa zea*, which are morphologically similar but differ in host plant preference and insecticide resistance (Pearce et al, 2017). Briefly, they assembled and annotated the genomes of *H. armigera* and *H. zea* and made phylogenetic and transcriptome analyses for gene families involved in detoxification, digestion and chemosensory functions, to explore the genomic basis of their differences in host range and insecticide resistance. The results indicated that *H. armigera* has 50 extra gustatory receptor genes and several more detoxification genes, which could explain why *H. armigera* is thought to be more polyphagous than *H. zea* (Pearce et al, 2017). Yet due to the lack of information about polymorphism within species, comparative genomics cannot be used to comprehensively address questions in the studies of intraspecific nucleotide and structural variations, and population studies within species are needed.

1.4 Lepidopteran sexual communication systems

The chemical ecology, courtship and mate-recognition behaviors, and molecular genetic machinery associated with lepidopteran pheromone-based sexual communication systems has been studied for decades. Pheromones were originally defined as chemical compounds that are secreted or excreted by one member of a species to elicit a specific reaction by other members of the same species. Unlike hormones, these chemicals are secreted outside of the body for communication between individuals (Karlson and Luscher, 1959). Sex pheromones are pheromones

released by one sex to attract members of an opposite sex, resulting in the location of the emitter and subsequent mating. They exist in almost all insect orders, including Lepidoptera, Trichoptera, Coleoptera, Heteroptera. Lepidopteran pheromones were the first to be widely studied (Sonenshine, 2016).

In moths, sex pheromones are released into the air by females from their pheromone glands (PGs). Males with specific ORs can behaviorally respond to calling females upon binding of the pheromone components. These males are immediately attracted to females, leading to upwind flight (Percy-Cunningham and MacDonald, 1987). Pheromones are potent attractants, even at low doses and over long distances. For example, males of the emperor moth can detect the sex pheromones of conspecific females up to 11 km away (Regnier and Law, 1968). However, not all pheromone plumes reach 11 km. Distances over which pheromone communication can occur varies among species from a few meters to more than 10 km, with moth groups such as the Saturniidae and Lymantriinae most likely to represent the upper end (Carde, 2016).

Female pheromones are a complex blend of chemical compounds, where the exact combinations of compounds and their ratios make up a species-specific blend, to which the conspecific male responds. Reproductive isolation can be achieved with such sex pheromone specificity (Roelofs and Comeau, 1969). Because of their species specificity, the chemical makeup of sex pheromones can also be used to distinguish between unique taxa. For example, two tortricid species, *Archips mortuanus* and *Archips argyrospilus* were first considered as one species by Powell, because the two forms have similar morphological adult characters, the same seasonal

cycles and hosts, and indistinguishable egg masses and young larvae. But they can be separated in their final larval instar (Powell, 1964). Later studies of sex pheromone specificity showed that these two forms are actually two species. They were attracted by different pheromone extracts, so the only isolating factor would be sex pheromone specificity (Roelofs and Comeau, 1969). Chemical studies then revealed that different mixtures were used by these two species. Male *A. mortuanus* were maximally attracted to the sex pheromone as a blend of Z11-14:OAc, Z9-14:OAc, E11-14:OAc, 12:OAc with a 90:1:10:200 ratio, while male *A. argyrospilus* were maximally attracted to a blend of the same components, but in a ratio of 60:40:4:200 (Carde et al, 1977).

Sex pheromone specificity is one potential factor that accelerated the evolution of Lepidoptera and produced the rich diversity of species seen today (Lofstedt, 1993). Multiple hypotheses have been proposed to describe how speciation occurs among Lepidoptera with pheromone-based sexual communication systems, including but not limited to the signal-response coevolution hypothesis, and the asymmetric tracking hypothesis. In the signal-response coevolution scenario, the evolution of moth mating systems requires a coordinated signal and response, leading to strong purifying selection on both the signaler and responder (Phelan, 1992; Lofstedt, 1993). In the process, a mutation occurring in the female, resulted in the production of a new female signal. Then the surviving female mutant mates with a male having mutations that allowed him to accept her pheromone signal, and they produce offspring with both divergent signaling and perception. This is difficult to imagine because the evolution of a new pheromone signal-response system requires a

coordinated change of both the signal sender and the receiver simultaneously, especially when the genes controlling female pheromone production and male responses are generally on different chromosomes (Roelofs et al, 1987; Lofstedt et al, 1989).

To address this conceptual problem, Phelan proposed another hypothesis to explain the evolution of sex pheromones, known as the asymmetric tracking theory. He argued that in most moth mating systems, selection will be largely unidirectional between signal sender and receiver (Phelan, 1997). In general, individual variation in female chemical signals is usually narrower than variation in the male response (Svensson, 1996). Changes in female pheromone blends require that corresponding males have broader responses to both wild-type and mutant female pheromones, resulting in a male population with a wide breadth in pheromone responses (Phelan, 1992).

Quantifying variation in male responses to the signals of females is labor intensive and time consuming. But molecular data from the odorant receptors (ORs) involved in male pheromone perception could be used to test hypotheses about how divergent signaling systems evolve in insects, ultimately leading to reproductive isolation and speciation. Under the signal-response coevolution scenario, these genes in both the males and females should be under strong purifying selection; while under the asymmetric tracking scenario, we would expect to see greater genetic variation in the ORs, allowing for a wide response window in males. The genetic variation may involve nucleotide diversity (polymorphism) among individuals, functionally redundant gene copies created by gene duplication (e.g. two or more ORs can detect

the same pheromone component) and neofunctionalization (e.g., one copy is able to detect a mutant pheromone component). Patterns of genetic variation at these chemosensory genes may provide important insights into the evolution of pheromone communication systems and speciation in Lepidoptera.

1.5 Diversification and divergence in the Heliothinae

My research focused on two lepidopteran species, *Chloridea (Heliothis) virescens* and *Chloridea (Heliothis) subflexa*. They are closely related and occur sympatrically throughout North and South America but differ in their host plant preferences. *Chloridea virescens*, also known as tobacco budworm, is an important pest of many crops in the United States, attacking cotton, flax, soybean, and tobacco in their larval stages (Sheck and Gould, 1993). *Chloridea subflexa* is a specialist, however, only feeding on a few species of plants in the genus *Physalis* (Laster et al, 1982; Groot et al, 2009). While *C. virescens* is not known to feed on *Physalis* in nature, their larvae can accept *Physalis angulata* under lab conditions. Moreover, males and females of *C. virescens* and *C. subflexa* are not attracted to each other in the wild, and hybridization is not known to occur in nature, yet they can be hybridized for genetic studies in the laboratory (Teal et al, 1995). Laboratory studies have suggested that the survival rate of *C. virescens* larvae using *P. angulata* fruit was ~10 times lower than that of *C. subflexa* larvae (Oppenheim et al, 2012). Further, recent studies of laboratory crosses found that the genetic architecture of host plant adaptation in *C. virescens* was complex. These studies found that the genetic loci under selection spread across half of the 31 chromosomes and were linked to multiple

traits (Oppenheim et al, 2018). Furthermore, the use of *Physalis* by *C. subflexa* also depended on many loci of small effect distributed throughout the genome (Oppenheim et al, 2012).

1.6 Research aims

The genetic basis of host plant adaptation and divergent sexual communication remains largely unknown among members of the Heliothinae. Application of comparative and evolutionary genomic approaches to Heliothine chemosensory gene families would provide new opportunities to fill these knowledge gaps. My dissertation focuses on the development of genomic tools to investigate the genetic mechanisms of host plant and mate recognition in Heliothine moths, and the application of these tools to examine inter- and intraspecific diversity of chemosensory genes of Heliothine moths.

In chapter 2, I produced an updated, highly contiguous *C. virescens* assembly and a novel *C. subflexa* Illumina short read assembly that could be used in comparative genomic studies. Short read assemblies present challenges for production of highly contiguous assemblies. Therefore, as part of this chapter, I examined the feasibility of reference-assisted assembly, an approach that leverages existing high quality genomic resources for genome improvement in closely related taxa and applied it to our Heliothine genomes. In chapter 3, I manually curated two chemosensory gene families, the odorant receptors (ORs) and odorant binding proteins (OBPs), that play crucial roles in odorant detection in *C. virescens*. Phylogenetic analyses were made to study their evolutionary divergence among

Heliothine moths. In chapter 4, I used field-collected *C. virescens* and *C. subflexa* to examine population-level genetic variation in their pheromone receptors (PRs) and to test hypotheses about the evolution of male pheromone responses to female chemical signals. As part of this chapter, I also studied host plant use of *C. virescens* and *C. subflexa* by comparing their larval densities and infestation rates in different host plants across the summer season. This was important for understanding where and when these species could be collected in the state of Maryland. Using these field-collected moths, I then inferred the strength and type of selection imposed on the pheromone receptor genes of each species and tested whether the genetic variation was associated with host plant use. The new genomic tools and the analyses of chemosensory gene families described in this dissertation will provide new opportunities for future investigation into the genetic mechanisms underlying host plant adaptation and sextual communication in Lepidoptera.

Chapter 2: Validation of reference-assisted assembly using existing and novel Heliothine genomes

2.1 Introduction

Genome assemblies serve as indispensable tools in modern entomological research. Early generation of genomic resources focused on production of annotated chromosome-scale assemblies in a small number of model insects (e.g. *Drosophila melanogaster*, Adams et al, 2000; *Anopheles gambiae*, Holt et al, 2002; *Bombyx mori*, Xia et al, 2004), which established the platform for modern insect molecular genetics. Insects show extraordinary morphological, physiological, and behavioral diversity, however. To reveal the genetic basis for this diversity and the evolutionary processes that shaped it, the Insect 5000 Genomes Initiative proposed sequencing and analysis of non-model species across the arthropod phylogeny (i5K Consortium, 2013). The primary goal was to generate thousands of publicly available reference genomes representative of diverse insect taxa. By July of 2020, 20 years after the publication of the *D. melanogaster* genome (Adams et al. 2000), 497 insect species had completed assemblies.

Low-cost short read sequencing enabled this proliferation of insect genome assemblies (Richards and Murali, 2015). Detection of polymorphisms within or between species, measurement of genome-wide transcriptional activity, discovery of gene order and duplication, as well as identification of DNA-protein interactions have all been made possible by mapping high throughput sequencing reads to these

important resources (Feuk et al, 2005; Xia et al, 2009; Ekblom and Wolf, 2014; Groot et al, 2016). Despite this progress, insect assemblies generated over this 20-year period vary in quality (see Appendix Table 2.1, Supplemental Data S1; sourced from National Center for Biotechnology Information (NCBI) and i5k platform July 16th 2020). Metrics of assembly quality include contiguity, completeness, and correctness, all of which are maximized in highly accurate assemblies (Thrash et al, 2020). It is known that insect genome assembly quality can be impacted by the laboratory, sequencing, and bioinformatic approaches used, as well as biological features of the organism (Li et al., 2019; Richards and Murali, 2015). The small physical size of some insects may limit the amount of genetic material available for sequencing and require pooling of individuals to obtain sufficient DNA. In diploid insects of moderate genome size, naturally high heterozygosity, the presence of repetitive DNA, and polymorphisms resulting from pooled sequencing can result in fragmentation of chromosomes into short scaffolds. In some cases, fragmentation of chromosomes can result in fragmentation of genes and reduced genome completeness (Ellis et al, 2021; Thrash et al, 2020).

Recently, long-read sequencing and assembly technology coupled with Hi-C has helped to overcome some of these challenges (Childers et al, 2021). Some insects have genomes of significant size (*e.g.* 5 Gb for bristletails and 7 Gb for grasshoppers), however, and large numbers of raw sequencing reads are needed to adequately cover the target species genome. This makes the use of long- and even short-read technology costly, especially for under-resourced insect taxa. A consequence of these fragmentation-, coverage-, and cost-related challenges is a lack

of genomic resources for parts of the arthropod phylogeny, particularly for taxonomic groups that are not of public health or agricultural importance. Another open question is how to proceed with hundreds of previously generated arthropod genomes of modest contiguity and completeness: Can we continue to improve on these genomic resources, or do all require resequencing and *de novo* assembly?

One approach to improving upon these already existing genomic resources is known as "reference-assisted assembly" (Lischer et al, 2017). The approach makes use of synteny between reference and target genomes to bioinformatically enhance the target genome contiguity and completeness. The risk is that differences in gene order and physical linkage along the chromosomes of reference and target genomes could result in mis-assemblies, thus lowering the genome's correctness. Newer approaches allow multiple references and order and orient short scaffolds by considering phylogenetic relationships to more conservatively identify true syntenic blocks. Reference-assisted approaches were initially used to assemble bacterial genomes sequenced by Illumina short reads (Kolmogorov et al, 2014; Gopinath et al, 2018) and genomes of different breeding lines of the plant, *Arabidopsis thaliana* (Schneeberger et al, 2011). More recently, algorithm modifications have been made to enable the use of closely related species (Bao et al, 2014; Lischer et al, 2017; Kolmogorov et al, 2018; Alonge et al, 2019).

Synteny-based approaches have been used on only a few insect species with more comprehensive genomic resources. For example, a gene synteny-based approach that used 21 *Anopheles* mosquito assemblies and the topology of the species phylogenetic tree was recently applied to super-scaffold the assemblies of important

mosquito vectors of disease (Waterhouse et al, 2020). In the process of super-scaffolding, short scaffolds were joined together into very large scaffolds. The contiguities of published draft genome assemblies for 16 *Heliconius* species were also improved by reference-assisted scaffolding, resulting in 95.7–99.9% of their genomes being anchored to chromosomes (Seixas et al, 2021). What has been less considered, however, is the accuracy and cost-benefit of using a reference-assisted approach. Given the challenges of producing continuous Illumina-only assemblies for insect species, and the costs of chromosome-scale assemblies using PacBio combined with Hi-C, we assessed whether and how a reference-assisted approach can improve low-cost Illumina assemblies. Lepidopteran species are a good test case for examining the feasibility of this approach because of their reported well-conserved macrosynteny between distantly related taxa, such as *B. mori* and *Heliconius* species (~ 100 Mya; Pringle et al, 2007). Therefore, our work explores the value proposition of reference-based assembly by using 4 non-model insect species in the agriculturally important super-pest clade Heliothinae, part of the family Noctuidae.

Focusing on the Heliothinae is strategic because it represents a reality for most highly diversified clades that include economically, ecologically, and evolutionarily important species. The insect family Noctuidae has >25,000 species (Rabieh 2018), but only 11 have genome assemblies. Of these, three are historically or currently important agricultural pests from the subfamily Heliothinae: *Chloridea* (*Heliothis) virescens* (Fritz et al, 2018), *Helicoverpa zea* and *H. armigera* (Pearce et al, 2017). Another North American member of the subfamily, *C. (Heliothis) subflexa*, is not a pest, and few genomic resources exist for this species. Establishment of these

resources is crucial, however, for understanding diversification of this pest lineage, as well as the evolution of pest traits. *C. subflexa* is a narrow specialist, feeding only on a few species of plants in the genus *Physalis* during its larval stages. Numerous studies have described *C. subflexa* host choice and sexual communication relative to the closely related polyphagous pest, *C. virescens* (Baker et al, 2004; Groot et al, 2005; Groot et al, 2011; Oppenheim et al, 2012; Oppenheim et al, 2018). Comparison of the *C. subflexa* genome to other Heliothine genome assemblies will offer future opportunities to identify mechanisms of sex pheromone production and detection, which underlie divergence from its pestiferous relatives, and could reveal evolutionary changes required for novel host plant detection and use.

Here, we updated the *C. virescens* genome using a Dovetail Genomics approach® and produced a short-read *C. subflexa* assembly to both expand and improve upon existing Heliothine genomic resources. We used these and other Heliothine genomes available on NCBI to examine the degree to which reference-assisted assembly increases contiguity and completeness while maintaining genome correctness (Kolmogorov et al, 2018). These Heliothine genomes were generated by Illumina short reads and are qualitatively similar to many of the publicly available insect genomes generated in the past 20 years, making them an excellent test case for reference-assisted assembly.

Our workflow first involved extending the published *C. virescens* assembly (N50 = 102,124 bp) using the two published *Helicoverpa* genomes (*H. armigera* and *H. zea*) and the new *C. subflexa* assembly as references. Reference-based *C. virescens* re-assemblies showed significantly increased contiguity and completeness. We then

measured the accuracy of this improvement strategy by comparing our reference-assisted assemblies against our new *C. virescens* Dovetail assembly which used Chicago® and Hi-C approaches for super-scaffolding. Comparison of the *C. virescens* reference-assisted re-assemblies to the Dovetail assembly validated the gene order correctness, and thus our approach. To further test the effectiveness of using reference-assisted assembly for low-budget genome improvement, we applied it to our new *C. subflexa* short-read assembly and the two fragmented *Helicoverpa* assemblies for super-scaffolding using the highly contiguous Dovetail *C. virescens* assembly. We conclude by placing the quality of the super-scaffolded Heliothine genomes into the context of other i5K genomes, and we discuss the potential and pitfalls of reference-assisted assembly considering our results. Our work adds to a growing body of literature indicating that reference-assisted assembly is an economical option for genome improvement (Bao et al, 2014; Lischer et al, 2017; Kolmogorov et al, 2018; Alonge et al, 2019).

2.2 Materials and methods

2.2.1 Insect genome assemblies

A table that summarized all of the arthropod genome projects until Oct 30, 2019 archived in an International Nucleotide Sequence Database Collaboration (INSDC) was downloaded from i5k website (http://i5k.github.io/arthropod_genomes_at_ncbi). Insect order, family, scientific name and scaffold N50 were extracted from the table for each arthropod genome (Supplemental Data S1). Information for insect genome assemblies completed from

Oct 30, 2019 to July 16, 2020 was summarized according to NCBI records (https://www.ncbi.nlm.nih.gov/assembly/organism/50557/all/). To examine the quality of lepidopteran assemblies, all 87 available lepidopteran genomes were downloaded from NCBI, and their completeness was assessed by BUSCO v4.0.6 (Seppey et al, 2019) using conserved gene sets of insecta_odb10 (Supplemental Data S2).

2.2.2 Genome size estimation

To accurately estimate the genome sizes of *C. virescens* and *C. subflexa*, two strategies were applied. First, *C. virescens* Illumina PE library SRR5463746 and one newly generated Illumina PE library of *C. subflexa* were used to perform computational estimation, by K-mer occurrence distribution (K = 19). Jellyfish (version 2.2.10; Marcais and Kingsford, 2011) was used to count the occurrence of K-mer and GenomeScope (Vurture et al, 2017) was used to estimate the genome size, heterozygosity and repeat content.

Second, their genome sizes were determined following methods in Johnston et al, (2019) using flow cytometry (FACSCanto II, the MPRI flow cytometry core, University of Maryland, College Park). The head of each frozen adult sample was placed in ice-cold Galbraith buffer, along with the head of an adult *Drosophila melanogaster* (1C = 175 Mb). Combined heads of the unknown sample and standard were ground using ~15 strokes of the "A" pestle in a 2 ml Dounce tissue grinder and filtered through a 20 μm nylon mesh. The DNA in the nuclei released by grinding was stained for 10 h under dark refrigeration with 50 μg/ml propidium iodide. The mean red PI fluorescence of stained nuclei was quantified, and the total quantity of

DNA in the sample was calculated by this formula: PI (sample)/PI (standard) * 175 Mb. To increase precision, the genome sizes for each sex of each species were estimated as the average of three biological replicates, and the standard deviation was also calculated.

2.2.3 Dovetail *C. virescens* assembly

Our Dovetail assembly was provided by Dovetail Genomics Company (https://dovetailgenomics.com). Illumina short read sequences from the original *C. virescens* genome (Fritz et al, 2018) were assembled into contigs, which were then scaffolded using Dovetail's proprietary approach. One pupa was used in Chicago® library preparation, which enables the identification and subsequent breaking of mis-joins, and then rejoins the fragments in the correct order and orientation giving longer and more accurate contiguity. Another pupa was used in Dovetail™ Hi-C sequencing library preparation. Short read assemblies were converted to ultra-long scaffolds by adding Hi-C, an approach that uses chromatin interaction information to group and linearly organize sequences into chromosomes. The pupae were provided by Benzon Research (https://www.benzonresearch.com/species-detail). The *C. virescens* lineage was originally collected and colonized at the USDA facility in Stoneville, Mississippi, and it has been in culture for >30 years.

2.2.4 *C. subflexa* Illumina library generation

C. subflexa collected in 1997 from Orangeburg County, South Carolina, and laboratory-reared at North Carolina State University were sib-mated for four generations. Genomic DNA was isolated from one male progeny from this fourth

generation of inbreeding using a Qiagen DNeasy kit (Qiagen, Valencia, CA, USA) using a modified mouse tail protocol. Isolated DNA was submitted to the North Carolina State University Genomic Sciences Laboratory (NCSU-GSL; Raleigh, NC, USA) for Illumina TruSeq Library preparation and sequencing on a single lane of the Illumina HiSeq 2500 in rapid run mode. Genomic DNA from a second outbred male pupa of the same colony was submitted to the Michigan State University Research and Technology Support Facility (East Lansing, MI, USA) for Illumina 2-kb and 5-kb mate pair (MP) library preparation. The MP library was prepared using an Illumina Nextera Mate Pair Sample Preparation Kit according to standard protocol, and validated using a Qubit dsDNA assay, Caliper LabChipGX (PerkinElmer, Waltham, MA, USA) and Kapa Library Quantification qPCR for Illumina Libraries. The library was sequenced on one lane of an Illumina HiSeq 2500 at the NCSU-GSL.

2.2.5 Evaluation of the novel *C. subflexa* short read assembly

We checked the quality of the Illumina reads using FastQC (Babraham Bioinformatics Cambridge, UK), and removed the Truseq adapter and Nextera transposon sequences from the reads using cutadapt v. 1.16 (Martin, 2011). The initial short read assembly was conducted using SOAPdenovo2 v. 2.04 (Luo et al, 2012), which uses a *de Bruijn* graphing approach to construct contigs. Twenty seven assemblies with both different k-mer size settings (17–127) and merge-level (M) settings (0 to 3) were generated to examine which parameter setting produced the most contiguous assembly. Nine assemblies with the highest scaffold N50s, produced with different combinations of parameter settings (k = 39, 41, 43 and M = 1, 2, 3) were selected for further examination of completeness by BUSCO v4.0.6 analysis

(Seppey et al, 2019). BUSCO groups in the Insecta lineage were used to measure the completeness of the 9 assemblies, and then the best assembly having both high contiguity and completeness was used for downstream analysis.

To resolve redundant haplotypes in the *C. subflexa* genome assembly, we used HaploMerger2 (version 20,180,603; Huang et al, 2017) according to standard protocols in the software documentation. Our new best *C. subflexa* assembly was repeatmasked with windowmasker (Morgulis et al, 2005) in the HaploMerger2 package. Only three modules were used: misjoin detection and removal (with the program hm.batchA), creating an initial haploid assembly, refining its single alleles, and creating a final haploid assembly (with hm.batchB), and removing tandem repeats from the haploid assembly (with hm.batchD). Then, a BUSCO analysis was rerun to determine whether processing by Haplomerger2 impacted the completeness by removing important gene-containing regions.

Upon confirming that Haplomerger2 did not impact genome completeness, we further scaffolded the genome assembly by previously published *C. subflexa* RNA-seq data using L_RNA_scaffolder (Xue et al, 2013). Briefly, the method identifies 'guide' transcript exons that are anchored to different scaffolds, by aligning transcriptome reads to genomic fragments. Scaffolding paths are constructed by walking the optimal connections of the anchored scaffolds, based on the number of transcripts mapped to them. Publicly available RNA-seq datasets of *C. subflexa* from different tissues including gut, thorax, male antennae and female antennae (Barthel et al, 2016; SRA accession numbers ERR738599, ERR738600, ERR738601, ERR738602, SRR6417881, SRR6417879, SRR6417883) were used to further extend

the scaffolds. For each dataset, the RNA-seq data were assembled into cDNAs with Trinity 2.8.3 (Grabherr et al, 2011) in *ab initio* form, and the new assembled transcriptomes were subsequently used to guide iterative scaffolding of the *C. subflexa* genome. Scaffolds smaller than 2 kb were filtered out. The assembly was reanalyzed in BUSCO to see the improvements in completeness.

2.2.6 Validation of the reference-assisted assembly using *C. virescens*

Ragout (version 2.2) is a software that can be applied to reconstruct fragmented genomes using syntenic and phylogenetic information from close relatives (Kolmogorov et al, 2018). To test whether this approach could be used to improve the quality of fragmented insect genome assemblies, the published *C. virescens* assembly (GenBank assembly accession: GCA_002382865.1; Fritz et al, 2018) that is similarly fragmented was used as the target assembly. We further scaffolded the published *C. virescens* assembly by Ragout 2.2 using the genomes of *C. subflexa*, *H. armigera* (GenBank assembly accession: GCA_002156985.1; Pearce et al, 2017) and *H. zea* (GenBank assembly accession: GCA_002150865.1; Pearce et al, 2017) as references. Whole-genome alignment of the four Heliothine genomes was made by SibeliaZ 1.1.0 (Ilia and Medvedev, 2019), a fast whole-genome aligner based on *de Bruijn* graphs. The resulting MAF format file was used as an input of Ragout 2.2 for scaffold extension, and different parameters "—solid-scaffolds True" and "—solid-scaffolds False" were used respectively. With the parameter "—solid-scaffolds" as false, the software was able to detect the chimeric module and break misjoined contigs that can cause large structural errors.

To evaluate and compare the extended assemblies, mapping statistics were measured by mapping the Illumina PE library SRR5463746 back to the published *C. virescens* assembly, the *C. virescens* Dovetail assembly and the two Ragout assemblies. The raw reads were randomly subset to 5 datasets, each containing 10% of the original data. They were mapped to the assemblies using Bowtie 2, and the mean mapping rate and standard deviation of each dataset were calculated. In addition, we mapped back ddRAD-seq reads. DdRAD-seq reads downloaded from Dryad digital data repository doi:https://doi.org/10.5061/dryad.4k40j were merged, filter-trimmed (see methods described in Fritz et al, 2018) and genome-aligned using the methods described above. These reads were generated from 5 *C. virescens* males collected by pheromone-baited light trap in Bossier Parrish, Louisiana, in 2012. N50 values were measured as a statistical indicator of contiguity using the assembly_stats program (https://github.com/sanger-pathogens/assembly-stats). BUSCO v4.0.6 was used to assess completeness of genes from the insecta_odb10 and eukaryote_odb10 databases.

2.2.7 Synteny and gene ordering

Each *C. virescens* Ragout assembly was compared with the Dovetail assembly using global genome alignment and visualized by syntenic dotplot using MUMmer 4.0.0 (Marcais et al, 2018). Hagfish (https://github.com/mfiers/hagfish) was used to estimate the insertion size distributions of the PE and MP libraries. Before using Hagfish, the read pairs were aligned to the assemblies by Bowtie 2 (Langmead et al, 2012). Then Hagfish was used to generate histograms of the insertion size

distributions of the libraries and identify potential misjoins, where the read pairs align either too far apart from each other or too close together.

A dense linkage map of *C. virescens* was generated in previous studies (Fritz et al, 2016), and the marker sequence data are available from Dryad Digital Repository (doi: https://doi.org/10.5061/dryad.567v8.). All the markers were grouped into 33 linkage groups, two more than the expected 30 autosomes and one Z chromosome (Fritz et al, 2016). The data was used to further evaluate the improved assemblies, to see if the software mis-joined scaffolds located on different linkage groups. In detail, the marker sequences were aligned back to the published *C. virescens* assembly, and the scaffold that each marker was located on was identified. If the scaffolds were used in further joining and extension by Ragout 2.2, their linkage groups were identified, which allowed us to determine whether the software mis-joined scaffolds from different linkage groups together from Ragout 2.2 reports.

After confirming reference-assisted assembly did not create major changes to synteny and gene order in the *C. virescens* Ragout assemblies, the genomes of *H. armigera*, *H. zea* (GenBank assembly accession: GCA_002156985.1 and GCA_002150865.1; Pearce et al, 2017) and *C. virescens* Dovetail assembly were used as reference genomes in Ragout 2.2 to build a super-scaffolded *C. subflexa* genome.

2.2.8 Macrosynteny with *B. mori*

All the super-scaffolded assemblies, including *C. sub_solid, C. sub_unsolid, H. zea_solid, H. zea_unsolid, H. armigera_solid* and *H. armigera_unsolid*, were compared with the reference genome of *B. mori* (downloaded from SilkDB3.0

(https://silkdb.bioinfotoolkits.net/main/help/-1#h-6) using MUMmer 4.0.0 (Marcais et al, 2018). Global protein alignment was made by promer and visualized by syntenic dotplot using mummerplot. Maximal unique matches (MUMs) were depicted as dots: red indicates forward matches and blue indicates reverse matches.

2.3 Results and discussion

2.3.1 Status of insect genomes in 2020

We identified insect genome assemblies available from the NCBI and i5K databases and produced quality metrics for them, allowing for their comparison to our own assemblies (see below). Our final dataset included assemblies for 497 insect species spanning 18 insect orders. As of July 2020, most insect assemblies had poor contiguity (defined by the USDA-ARS Ag100Pest Initiative, Childers et al, 2021, as scaffold N50 < 10 Mbp; Figure 2.1a), likely due to the difficulty of resolving repetitive sequences when their lengths are longer than the sequencing reads, as well as high insect heterozygosity. More recently, integration of long PacBio and Nanopore reads with Illumina short reads has been used to resolve poor contiguity (see improved assemblies of *Aedes aegypti*, yellow fever mosquito, yellow fever mosquito, Matthews et al, 2018; *Trichoplusia ni*, cabbage looper, Fu et al, 2018; *Cydia pomonella* version 2, codling moth, Wan et al, 2019), but required additional sequencing of the focal taxon. If reference-assisted assembly can adequately improve existing short-read assemblies, resources could be used to produce novel, high quality assemblies of close relatives, rather than re-assembly of existing genomes. The result would be substantial improvement of existing genomes, as well as expansion of

genomic resources across the arthropod phylogeny. Such an expansion of resources is essential for a full ascertainment of the evolutionary relationships between organisms (*e.g.*, Zhang et al, 2000).

Among insect orders, the Lepidoptera had the third highest number of assembled genomes (*n* = 87 species; Appendix Table 2.1), yet draft assemblies were only available for 23 of 137 lepidopteran families (Appendix Table 2.2). Lepidopteran genomes are of relatively small size (78 Mb–1280 Mb), with most being *ca.* 400 Mb (*n* = 42), and their N50 values ranged from 713 to 22,146,069 bp (median of 499,611 bp; Figure 2.1b). Although dispersed repeat content can account for a substantial proportion of their genomes (Cabral-de-Mello et al, 2021), most of the 87 lepidopteran assemblies were relatively complete (range = 16.24%–98.83%, mean = 83.33%, median = 93.71%), according to BUSCO analysis (v. 4.0.6; Insecta_odb10; Seppey et al, 2019). Species from the Nymphalidae made up the greatest proportion of assembled lepidopteran genomes (*n* = 32), but many species were fragmented or had very low completeness (Figure 2.1c & d, Supplemental Data S2). Variation in completeness was 4 times greater in the Nymphalidae compared to other lepidopteran families, possibly resulting from support for specific experiments rather than intent to generate genomic resources. In contrast, only 11 of the >25,000 Noctuid species had genome projects. This revealed a need for additional genomic resources to expand knowledge of genome structure and gene content within this speciose group (Appendix Table 2.2).

2.3.2 Genome size estimation

A K-mer analysis was applied to estimate the genome sizes of *C. virescens* and *C. subflexa*, using publicly available Illumina PE library of *C. virescens* (SRA accession number: SRR5463746) and one newly generated Illumina PE library of *C. subflexa*, respectively. This revealed genome sizes of 348 Mb with 0.523% heterozygosity and 23.94% repeat content, and 392 Mb with 0.486% heterozygosity and 33.40% repeat content, respectively (K = 19; Figure 2.2a and 2.2b). Estimates of genome size were also made with flow cytometry and were 381.99 ± 2.17 Mb and 466.87 ± 7.93 Mb, for *C. virescens* and *C. subflexa*, respectively (Figure 2.2c, d, Appendix Table 2.3). Comparing K-mer and flow cytometry-based estimates of genome size revealed that K-mer analyses underestimated genome size. This likely resulted from K-mer occurrence distributions that do not perfectly fit the assumed single Poisson distribution due to the heterozygosity and repeat content of these species (Liu et al, 2013). When compared to other Heliothines, *C. subflexa* had the largest genome, followed by *H. assulta* (430 Mbp; Zhang et al, 2019), *H. armigera* (394 Mbp; Zhang et al, 2019), *C. virescens*, and *H. zea* (362 Mbp; Coates et al, 2017). Repetitive elements are likely responsible for the increased *C. subflexa* genome size, an hypothesis supported by the percent repeat content estimate from our K-mer analysis.

2.3.3 Dovetail improvement of the *C. virescens* assembly

An improved *C. virescens* assembly was produced by applying Chicago® sequencing and Dovetail™ Hi-C sequencing to group and linearize short contigs into 31 super-scaffolds and an additional 583 shorter, unplaced scaffolds (total unplaced

scaffold length: 15,720,849 bp). The largest 31 super-scaffolds, which correspond to the 31 *C. virescens* chromosomes, contain 96.11% of the total assembly size (388,286,154 bp of 404,007,003 bp), 37,386,834 Ns and 52,086 gaps. Compared to the *C. virescens* assembly available on NCBI (GenBank assembly accession: GCA_002382865.1; Fritz et al, 2018), the Dovetail assembly has 8212 fewer scaffolds, a scaffold N50 value 134 times higher, and 12 additional complete BUSCOs (Table 2.1).

2.3.4 Evaluation of the novel *C. subflexa* short read assembly

A total of 364,910,444 and 172,596,248 Illumina short reads from paired-end (PE) and mate-pair (MP) libraries were used to produce a *C. subflexa* assembly (Appendix Table 2.4), which comprised 7654 scaffolds with a scaffold N50 of 264,374 bp (mean size = 60,913 bp; range = 2000–1,469,202 bp). The total assembly length was 466,227,171 bp, close to the flow cytometry estimate 466.87 Mb and greater than other Heliothines even after merging alternative haplotypes of the same genetic locus. A BUSCO analysis indicated that 90.8% of the 1367 conserved insect genes were fully assembled, and 90.6% of conserved genes were assembled as single copies (Table 2.1). The contiguity of this *de novo* Illumina assembly was >36.78% of the Lepidopteran assemblies and the completeness was above 45.98% (Supplemental Data S2). Assembly accuracy was also examined for a region of the genome containing a cluster of four olfactory receptor genes: OR6, OR14, OR15, and OR16 (GenBank accession numbers HM751833.1, HM751835.1, HM751836.1, HM751837.1). This provided novel insight into assembly accuracy because it moved

beyond conserved genes to focus on a more rapidly evolving gene family. We chose these genes because the available genomic resources for *C. subflexa* were limited, and we wished to use olfactory receptor genes that had orthologues in all Heliothines investigated here. Manual inspection of alignments between the four *C. subflexa* partial cDNA sequences and our novel assembly indicated that ORs 14–16 were ordered correctly on Scaffold 656 and assembled with nearly 100% similarity to their partial cDNA sequences (Appendix Table 2.5). OR6 mapped to its own scaffold (Scaffold 489) but also showed nearly 100% similarity to its partial cDNA sequence (Appendix Table 2.5). We attribute the average mismatch rate of 0.256% across these four ORs to the high rate of within-species polymorphism in Lepidoptera.

2.3.5 Contiguity and completeness of the *C. virescens* reference-assisted assembly

We examined the feasibility of reference-assisted assembly for genome improvement, using *C. subflexa* (see above), *H. armigera* (GenBank assembly accession: GCA_002156985.1; Pearce et al, 2017) and *H. zea* (GenBank assembly accession: GCA_002150865.1; Pearce et al, 2017) to super-scaffold the publicly available *C. virescens* genome (GenBank assembly accession: GCA_002382865.1; Fritz et al. 2018). We used Ragout 2.2, which stitches shorter scaffolds together after analyzing syntenic relationships and possible genome rearrangements between 1 or more references and a target genome (Kolmogorov et al, 2019). Gaps can be added within the stitches if they are supported by the multi-genome alignment. Two reference-assisted assemblies were generated using the different chimeric module settings of Ragout 2.2. The “unsolid” option broke putatively misjoined scaffolds that cause structural errors, and then stitched together the “broken” scaffolds according to

the reference, while the "solid" option did not break putative misjoins and only stitched together existing scaffolds.

We initially examined assembly accuracy by quantifying and comparing mapping rates of raw reads from the single pupa used to generate the publicly available *C. virescens* assembly (SRA accession: SRR5463746), as well as publicly available Double Digest Restriction-site Associated DNA sequencing (ddRAD-seq) reads from wild field-collected samples (Fritz et al. 2018) to both Ragout assemblies, the original *C. virescens* assembly and the Dovetail assembly. Raw pupal reads mapped to all assemblies at a rate of *ca.* 87%, with little variation around the mean (Appendix Table 2.6). A similar pattern was observed for mapping of the filter-trimmed and merged ddRAD-seq data, but at an equivalently reduced rate across assemblies (*ca.* 55%, s.d. 0.2–0.3%, Appendix Table 2.6). Overall, the equivalent mapping rates for all 4 assemblies served as one indicator that reference-assisted assembly does not strongly reduce assembly accuracy. We attributed the overall reduction in ddRAD-seq mapping rates across assemblies to high rates of polymorphisms known to occur among *C. virescens* individuals (Fritz et al. 2016).

Compared to the original *C. virescens* assembly, we observed a modest increase in contiguity in the reference-assisted assemblies, as measured by N50 (Table 2.2). This was promising considering the relatively fragmented nature of the reference *Helicoverpa* and *C. subflexa* genomes. Yet the total sizes of the reference-assisted assemblies increased, mostly due to the addition of 23,978,162 and 24,466,305 Ns (*i.e.*, gaps). BUSCO analyses revealed that an additional 0.3% of the conserved single copy insect orthologs were complete in the reference-assisted

assemblies compared to the published *C. virescens* assembly, resulting from a reduction of fragmented genes (from 2.3% to ~1.9%; Table 2.1, Table 2.2). Similar trends were observed for single copy eukaryote orthologs. Sequence accuracy and order of ORs 6, and 14–16, located on scaffold NWSH01000007.1 in the publicly available *C. virescens* assembly, were also retained in both assemblies produced by Ragout 2.2.

2.3.6 Synteny and gene ordering for reference-assisted *C. virescens* assemblies

Both Ragout assemblies were compared with the Dovetail assembly to examine gene ordering, and the results indicated high synteny with very minor syntenic discontinuities (Figure 2.3). Hagfish (https://github.com/mfiers/hagfish) was used to further assess the quality of the extended assemblies by comparing the insert size distributions of mapped read pairs. Read pairs that align closer or farther apart than expected based on the known insert size distribution indicate potential mis assemblies. Alignment results indicated that most read pairs had inserts of *ca.* 480 bp, which was close to the expected 550 bp insert size-selection used in the library preparation (Appendix Figure 2.1). Moreover, all the assemblies had the same insert size distributions, and no substantial mis-joins, structural variants, or errors were identified in the original contigs. This same pattern was seen with the two mate-pair libraries (Appendix Figures 2.2 and 2.3), indicating that reference-assisted assembly did not generate substantial structural errors. Alignment of the mate pair library showed two peaks: one at *ca.* 5.7 kb, which was close to the expected mate-span of 5 kb, and a second less prominent peak at *ca.* 8 kb, which indicated that the mate pair library was contaminated with 8 kb mate-spans during library preparation.

We also examined whether Ragout mis-joined scaffolds from different chromosomes by comparing the reference-assisted *C. virescens* assemblies to a previously published linkage map (Fritz et al. 2016). Our expectation was that the Ragout-joined scaffolds should map to markers from a single linkage group in the previously published linkage map. Six hundred sixty-two linkage marker sequences in 33 linkage groups were downloaded from Dryad Digital Repository (doi: https://doi.org/10.5061/dryad.567v8, Fritz et al, 2016, Supplemental Data S3A). Among them, 535 markers were aligned to 463 scaffolds of the published *C. virescens* assembly (Supplemental Data S3B). Not all of these 463 scaffolds were super-scaffolded during the Ragout extension process. In the *C. virescens_solid* assembly, 336 of the 463 scaffolds were super-scaffolded, and only one linkage marker mapped to 91 of these, preventing further analysis. Of the remaining 245, placement of scaffolds in the Ragout-extended assembly indicated that 220 had been properly joined (accuracy 89.8%) and 25 were potential mis-joins (error rate 10.20%; Supplemental Data S3C). For the *C. virescens_unsolid* assembly, 235 scaffolds had multiple marker alignments. Of those, 217 had no detectable mis-assemblies (accuracy 92.34%), 18 had potential mis-joins (error rate 7.66%; Supplemental Data S3D). After considering the syntenic relationship between the Ragout assemblies with the Dovetail assembly, the insert size distributions for mapped PE/MP reads, and validation with our previously published linkage map, we determined that reference-assisted assembly improved the contiguity and completeness of the previous *C. virescens* assembly without significantly affecting its correctness. However, the

approach also increased the assembly size through gap opening and insertion of ambiguous nucleotides (Ns).

2.3.7 Application of Ragout to *C. subflexa* and *Helicoverpa* short read assemblies

To extend our Illumina short-read *C. subflexa* assembly, we applied Ragout 2.2 using the Dovetail *C. virescens* and two publicly available *Helicoverpa* assemblies as references. The improved *C. subflexa* genome assemblies were comprised of 6354 scaffolds with a total length of 515,453,367 bp (*C. sub_solid*, 44.54% is Ns), and 5125 scaffolds with a total length of 551,569,823 bp (*C. sub_unsolid*, 48.18% is Ns; Table 2.3). BUSCO completeness scores for the insecta_odb10 gene set reached *ca.* 91% in the reference-assisted *C. subflexa* assemblies (Table 2.3). Contiguity and completeness of the *H. zea* (GenBank assembly accession: GCA_002150865.1; Pearce et al, 2017) and *H. armigera* (GenBank assembly accession: GCA_002156985.1; Pearce et al, 2017) assemblies also improved (Figure 2.4). For *H. zea*, the numbers of scaffolds were reduced by half, there was an increase in the number of complete insect and eukaryote genes, and the N50 values were 30–44 times larger following Ragout super-scaffolding (Appendix Table 2.7). For *H. armigera*, the reference-assisted assemblies had up to 404 fewer scaffolds and larger contig N50 values (Appendix Table 2.8). When placed in context of other lepidopteran genomes, the contiguities of *C. sub_solid, C. sub_unsolid, H. zea_solid, H. zea_unsolid, H. armigera_solid* and *H. armigera_unsolid* assemblies were above 70.11%, 98.85%, 82.76%, 86.21%, 89.66% and 89.66% of the Lepidopteran assemblies, and the completenesses were above

35.63%, 36.78%, 73.56%, 71.26%, 78.16% and 74.71%, respectively (Supplemental Data S2).

2.3.8 Macrosynteny with *B. mori*

We visualized synteny between the super-scaffolded assemblies and the reference genome of *B. mori* (Figure 2.5). *B. mori* was chosen for the comparison because it has a chromosome-scale reference genome, and macrosynteny among Lepidoptera is well-established (Pringle et al, 2007). All super-scaffolded assemblies had good synteny with *B. mori* except *C. sub_solid* (Figure 2.5b). One possible reason is that *C. sub_solid* is more fragmented than other assemblies, and its shorter scaffolds with greater gene fragmentation make it harder to estimate synteny based on protein-coding sequences. Altogether, this suggests that Ragout 2.2 provides more accurate synteny and gene ordering when it is allowed to detect and break the original chimeric contigs (with –solid-scaffolds set to false).

2.3.9 Advantages and disadvantages of reference-assisted assembly

One of the earliest and most emphasized measures of assembly quality was contiguity, although it is well-known that it cannot stand alone to determine genome accuracy (*e.g.*, Hunt et al., 2013; Salzberg et al., 2012). Contiguous assemblies are instrumental tools for some types of genomic studies, including quantitative trait locus analysis and genome-wide association studies. A lack of contiguity can even be problematic for the study of specific genes if their exons are spread across multiple scaffolds. Our data suggest that reference-assisted assembly can significantly improve the contiguity of a fragmented short-read assembly, and in doing so, can also increase

the number of intact, single copy genes. This increased genome contiguity and completeness can add value to low budget genome sequencing projects, making them more useful for downstream applications. In the case of *C. virescens*, the resultant increase in completeness was modest, due to the relatively high completeness of the original assembly. For species where completeness is low due to genome fragmentation, reference-assisted assembly is likely to be even more beneficial, with some caveats (see below).

As pointed out by Thrash et al, 2020, an increase in contiguity produced through scaffold mis-joins reduces assembly accuracy. When we carefully examined our reference-assisted *C. virescens* assemblies for evidence of mis-joins, our results indicated that reference-assisted assembly does not increase contiguity at the expense of correctness. This may be due to the conserved synteny and gene order known to exist among lepidopteran species (Pringle et al, 2007). For example, *Heliconius* species diverged from each other about 10 million years ago (Kozak et al, 2015), yet their genomes remain highly co-linear (Davey et al, 2017; Seixas et al, 2021). Indeed, Heliothine species are even more closely related (Cho et al, 2008), and our study combined with others (Seixas et al, 2021) demonstrates that reference-assisted extension of fragmented assemblies should be effective for any species group where co-linearity is expected.

However, a key disadvantage revealed by our work is that reference-assisted assembly introduces gaps between scaffolds and fills them with Ns, which can inflate genome size. A more complex issue is the highly heterochromatic chromosomes (such as the Y/W), which are repeat-rich and poor in protein coding genes, and many

of them are missing even from high-quality assemblies (Chang et al, 2019). We speculate that such regions are also unlikely to be well served by reference-assisted assembly. In addition, reference-assisted assembly introduces gaps that contain repetitive regions that could play important roles in structural and functional processes (Gemmell et al, 2021).

A final issue is the presence of biases due to the choice of reference genomes. Our data demonstrate that reference-assisted assembly is feasible when executed with high quality reference genomes, especially when the reference taxon is closely related to the target taxon. As divergence between the reference and target taxa, or particular chromosomes in those taxa increase, increased gap opening and expansion of genome size may occur. In addition, errors in the reference assemblies used in the extension process may be propagated to the target assembly (Lischer et al, 2017). Therefore, we expect this approach is best applied to sequencing projects where resources are limited and high-quality assemblies of closely related taxa are available.

2.4 Conclusion

As a growing number of insect genomes are being sequenced, reference-assisted assembly provides an opportunity to use existing insect genomes of closely-related species to assist the *de novo* assembly of a new species, or *vice versa*. Considering the cost and time associated with experimental finishing or re-sequencing efforts, reference-assisted assembly provides a cheap and effective way to arrange and anchor short scaffolds into chromosomes, at least until further molecular data can be generated. Our work, which generated a novel *C. subflexa* short read

assembly and an updated, highly contiguous *C. virescens* assembly, also demonstrated that reference-assisted assembly has the potential to enhance genome contiguity and completeness, making it a promising approach to leveraging and improving upon available genomic resources for insects.

Tables: Chapter 2

Table 2.1. Metrics of quality for the *C. virescens* (Fritz et al, 2018), Dovetail *C. virescens*, and *C. subflexa* assemblies.

Statistics	***C. virescens***	**Dovetail *C. vir***	***C. subflexa***
Assembly size (Mb)	403.15	404.01	466.23
Estimated size (Mb)	382	382	467
Number of scaffolds	8,826	614	7,654
Max scaffold length (bp)	628,964	19,818,128	1,469,202
Min scaffold length (bp)	659	629	2000
Mean scaffold length (bp)	45,678	657,992	60,913
N50 (bp)	102,124	13,658,951	264,374
N90 (bp)	21,868	9,753,174	33,773
N_count (bp)	38,034,845	38,888,196	180,381,610
BUSCO insecta_odb10			
complete (C)	1222 (of 1367, 89.4%)	1234 (of 1367, 90.3%)	1241 (of 1367, 90.8%)
single copy (S)	1187 (86.8%)	1197 (87.6%)	1238 (90.6%)
duplicated (D)	35 (2.6%)	37 (2.7%)	3 (0.2%)
fragmented (F)	31 (2.3%)	21 (1.5%)	47 (3.4%)
missing (M)	114 (8.3%)	112 (8.2%)	79 (5.8%)
BUSCO eukaryota_odb10			
complete (C)	219 (of 255, 85.9%)	227 (of 255, 89.1%)	224 (of 255, 87.8%)
single copy (S)	216 (84.7%)	223 (87.5%)	224 (87.8%)
duplicated (D)	3 (1.2%)	4 (1.6%)	0 (0%)
fragmented (F)	11 (4.3%)	4 (1.6%)	19 (7.5%)
missing (M)	25 (9.8%)	24 (9.3%)	12 (4.7%)

Table 2.2. Metrics of quality for the Ragout improved assemblies of *C. virescens*.

Statistics	***C. virescens_solid***	***C. virescens_unsolid***
Assembly size (Mb)	427.13	427.62
Estimated size (Mb)	382	382
Number of scaffolds	6,432	6,482
Max scaffold length (bp)	9,332,071	10,482,267
Min scaffold length (bp)	629	142
Mean scaffold length (bp)	66,407	65,970
N50 (bp)	357,313	380,458
N90 (bp)	27,912	27,785
N_count (bp)	62,013,007	62,501,150
BUSCO insecta_odb10		
complete (C)	1226 (of 1367, 89.7%)	1225 (of 1367, 89.7%)
single copy (S)	1191 (87.1%)	1190 (87.1%)
duplicated (D)	35 (2.6%)	35 (2.6%)
fragmented (F)	25 (1.8%)	27 (2.0%)
missing (M)	116 (8.5%)	115 (8.3%)
BUSCO eukaryota_odb10		
complete (C)	223 (of 255, 87.5%)	222 (of 255, 87.1%)
single copy (S)	218 (85.5%)	217 (85.1%)
duplicated (D)	5 (2.0%)	5 (2.0%)
fragmented (F)	8 (3.1%)	8 (3.1%)
missing (M)	24 (9.4%)	25 (9.8%)

Table 2.3. Metrics of quality for the improved assemblies of *C. subflexa.*

Statistics	***C. sub_solid***	***C. sub_unsolid***
Assembly size (Mb)	515.45	551.57
Number of scaffolds	6,354	5,125
Max scaffold length (bp)	7,923,169	22,413,266
Min scaffold length (bp)	2,000	101
Mean scaffold length (bp)	81,123	107,623
N50 (bp)	1,335,006	16,413,121
N90 (bp)	47,640	73,821
N_count (bp)	229,607,806	265,724,262
BUSCO insecta_odb10		
complete (C)	1238 (of 1367, 90.5%)	1241 (of 1367, 90.8%)
single copy (S)	1236 (90.4%)	1238 (90.6%)
duplicated (D)	2 (0.1%)	3 (0.2%)
fragmented (F)	48 (3.5%)	44 (3.2%)
missing (M)	81 (6.0%)	82 (6.0%)
BUSCO eukaryota_odb10		
complete (C)	226 (of 255, 88.6%)	226 (of 255, 88.6%)
single copy (S)	225 (88.2%)	225 (88.2%)
duplicated (D)	1 (0.4%)	1 (0.4%)
fragmented (F)	17 (6.7%)	17 (6.7%)
missing (M)	12 (4.7%)	12 (4.7%)

Figures: Chapter 2

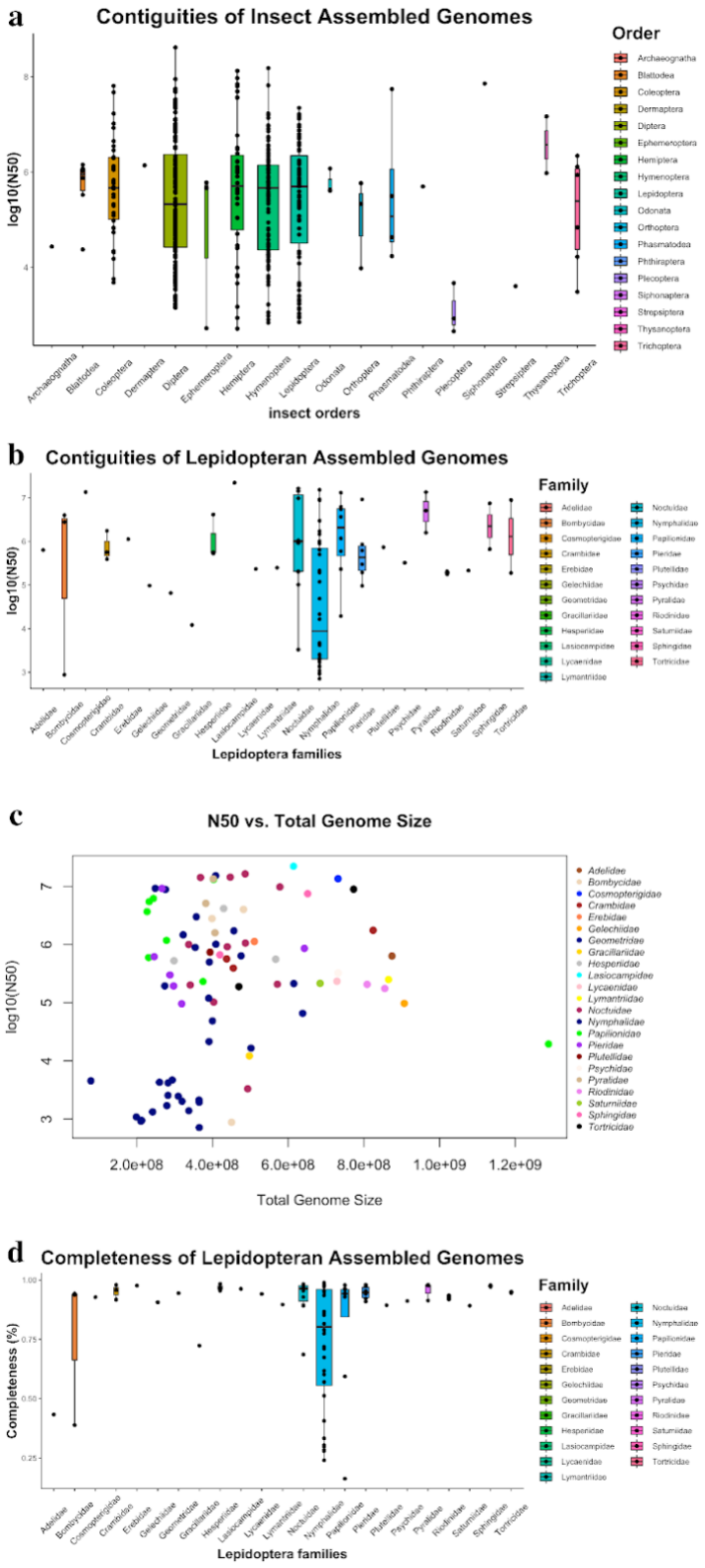
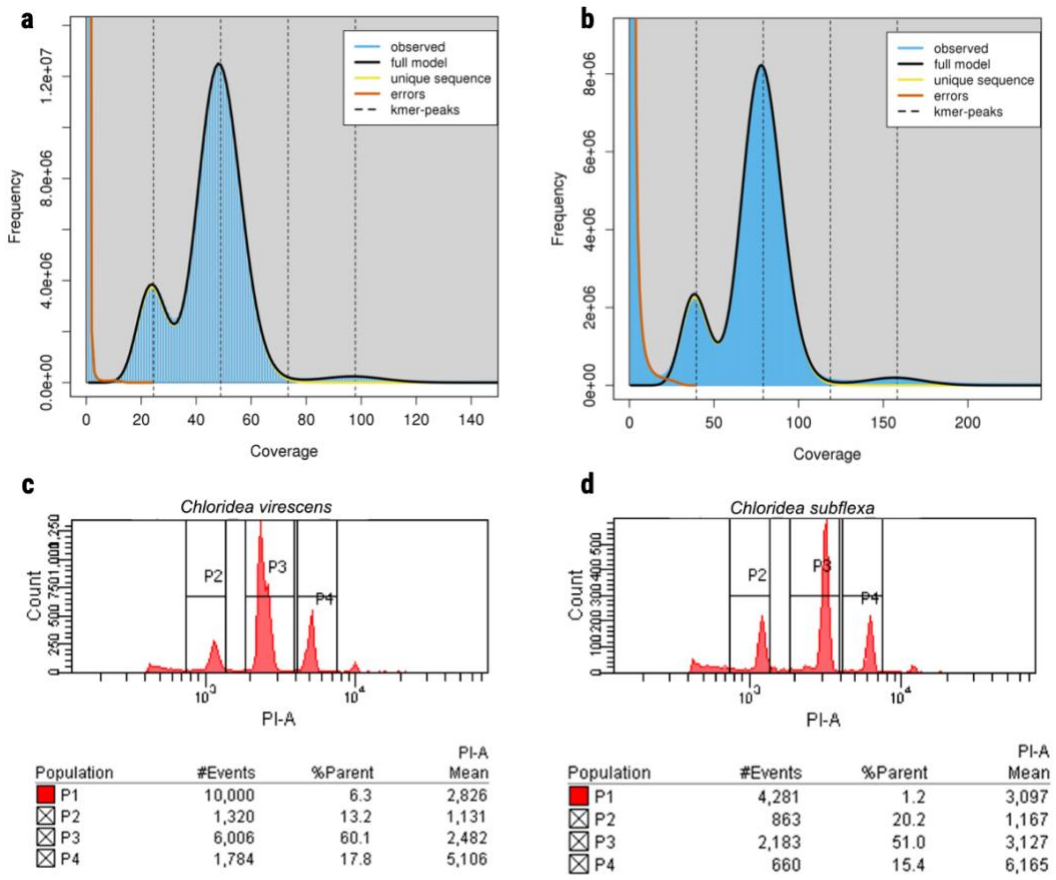


Figure 2.1. Contiguities of insect assembled genomes (a). Contiguities of Lepidoptera assembled genomes (b). Correlation of N50 vs. total genome size (c). Completeness of Lepidoptera assembled genomes (d).



Population	#Events	%Parent	PI-A Mean
P1	10,000	6.3	2,826
P2	1,320	13.2	1,131
P3	6,006	60.1	2,482
P4	1,784	17.8	5,106

Population	#Events	%Parent	PI-A Mean
P1	4,281	1.2	3,097
P2	863	20.2	1,167
P3	2,183	51.0	3,127
P4	660	15.4	6,165

Figure 2.2. *C. virescens* and *C. subflexa* genome size estimation. *C. virescens* (a) and *C. subflexa* (b) 19-mer occurrence distribution made by GenomeScope. *C. virescens* (c) and *C. subflexa* (d) genome size estimation using flow cytometry.

P2 is the control, *D. melanogaster*, P3 is the target species, *C. virescens* or *C. subflexa* and P4 is the mitotic cells with 2-fold greater DNA content.

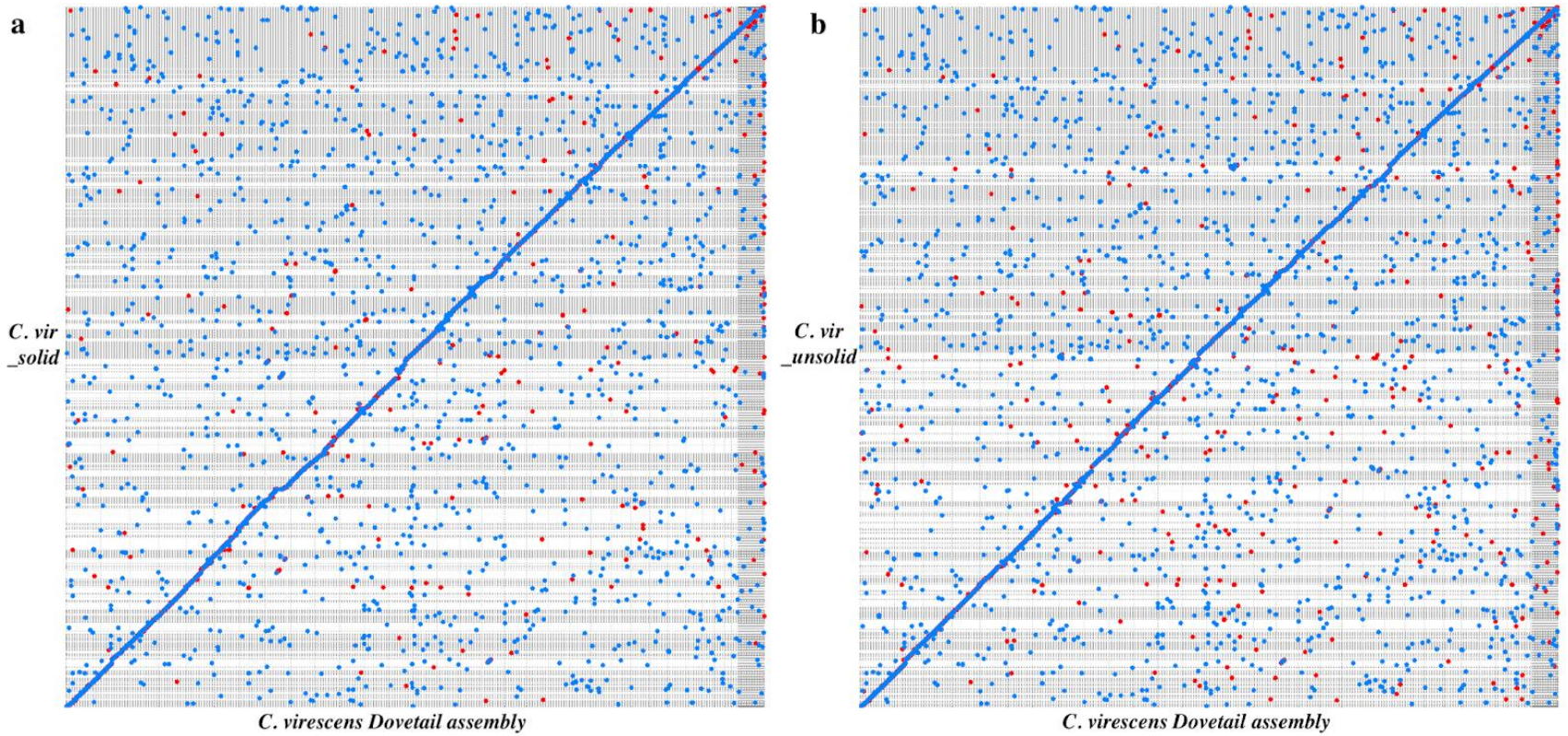


Figure 2.3. Dotplot comparison of the *C. virescens_solid* (a) and *C. virescens_unsolid* (b) assemblies to the *C. virescens* Dovetail assembly.

MUMmer 4.0.0 (Marcais et al, 2018) was used for alignment, and maximal unique matches (MUMs) are depicted as dots: red indicates forward matches; blue indicates reverse matches.

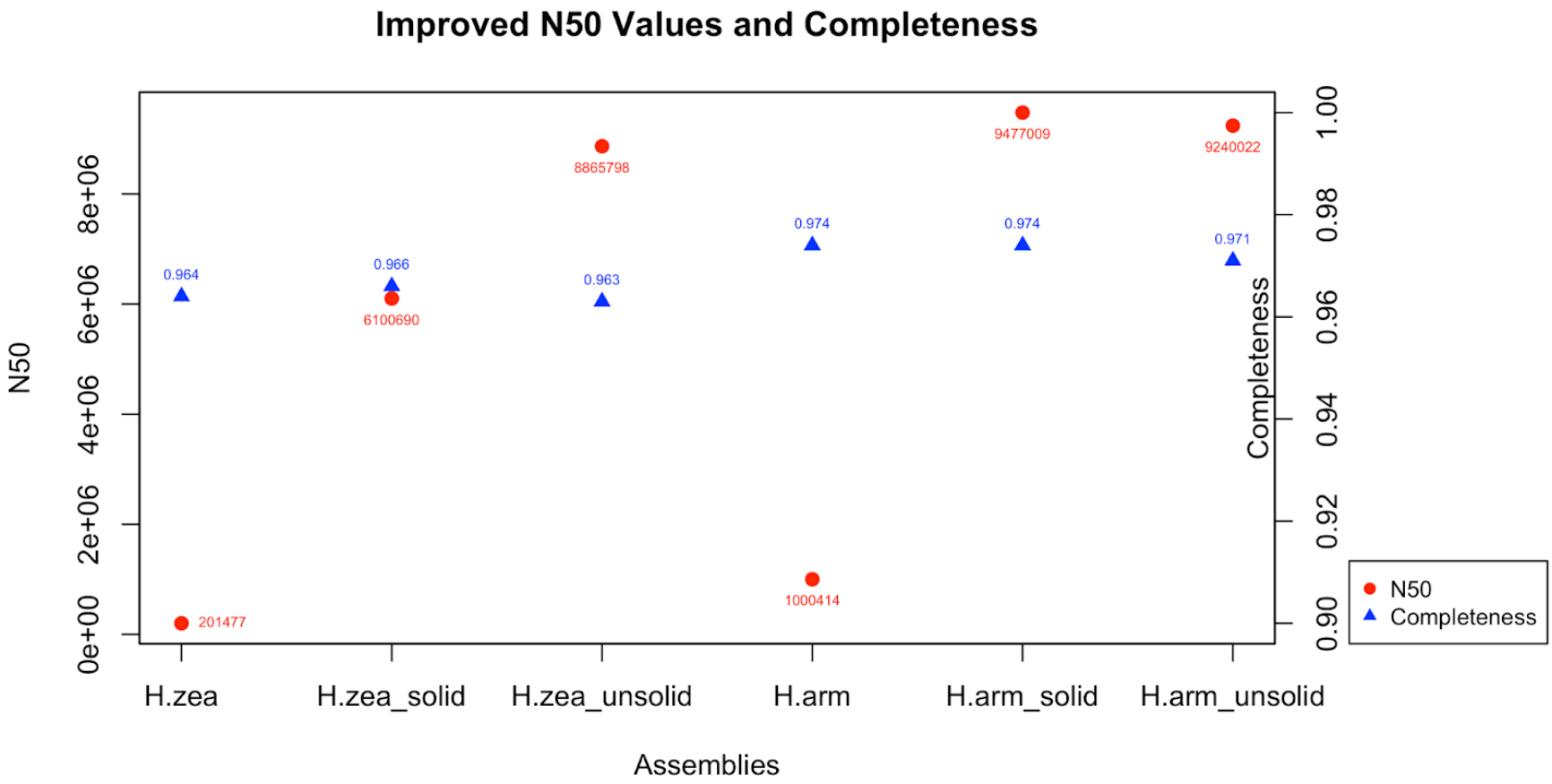


Figure 2.4. Contiguity and completeness changes of *H. zea* and *H. armigera* assemblies.

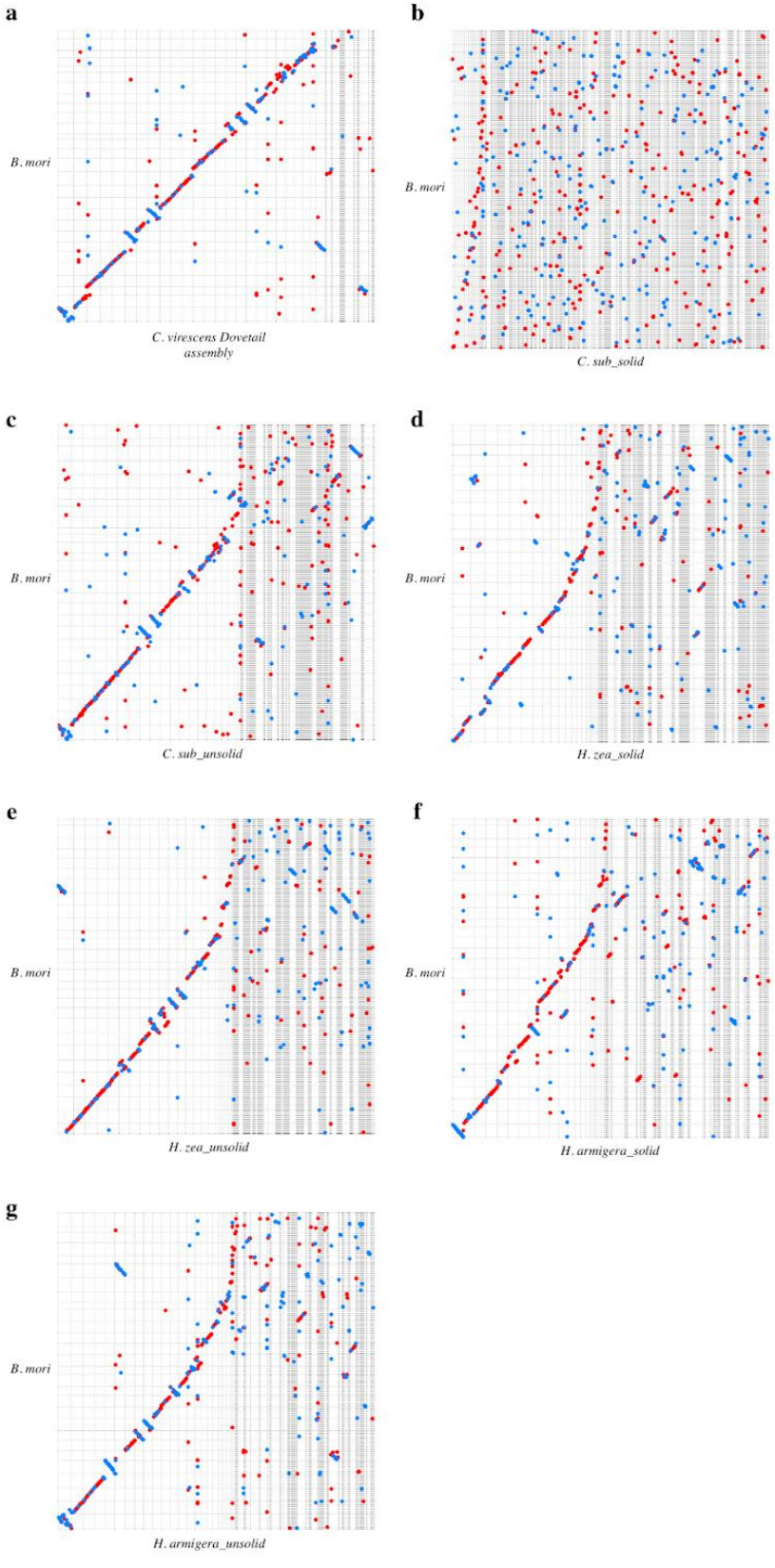

a
B. mori
C. virescens Dovetail assembly
b
B. mori
C. sub_solid
c
B. mori
C. sub_unsolid
d
B. mori
H. zea_solid
e
B. mori
H. zea_unsolid
f
B. mori
H. armigera_solid
g
B. mori
H. armigera_unsolid

Figure 2.5. Macrosynteny analysis with *B. mori*. Syntenic dotplots of the super-scaffolded assemblies *versus* the reference genome of *B. mori*. *C. virescens* Dovetail assembly (a). *C. sub_solid* assembly (b). *C. sub_unsolid* assembly (c). *H. zea_solid* assembly (d). *H. zea_unsolid* assembly (e). *H. armigera_solid* assembly (f). *H. armigera_unsolid* assembly (g).

Chapter 3: Evolutionary divergence of olfactory receptor (OR) and odorant binding protein (OBP) gene families among members of the Heliothinae

3.1. Introduction

The chemosensory system plays a crucial role in many insect behaviors, such as feeding, mating, finding oviposition sites and avoiding harmful factors such as predators (Leal et al, 2013). Accurate recognition and perception of certain odorants are critical to fitness, particularly pheromone components and host plant odorants (Zhang et al, 2015; Pelosi et al, 2018). In the process of sexual communication in Lepidoptera, sex pheromones are detected by olfactory sensory neurons (OSNs) housed in multiporous cuticular hairs on the antenna. The hydrophobic pheromone molecules are first absorbed by the cuticular surface, and then pass through pores of the olfactory sensillum. The molecules become water-soluble through binding to the odorant binding proteins (OBPs) in the lymph fluid, and they are carried to OSN dendrites, finally binding with pheromone receptors (PRs), special odorant receptors (ORs), on the membrane of OSN dendrites (Sakurai et al, 2014). The binding triggers the activation of ion channels and converts chemical signals into electrical signals.

ORs belong to the G-protein-coupled receptor (GPCR) family, all of which are transmembrane proteins containing seven domains. Membrane topology analysis of the insect ORs revealed that they have an intracellular N-terminus and an extracellular C-terminus, making them distinct from other GPCRs which have the C-

termini inside the cell (Smart et al., 2008; Tsitoura et al., 2010; Carraher et al, 2015). Insect ORs function as a heterodimer in OSN: a co-receptor Orco, previously known as OR83b (Vosshall and Hansson, 2011), which is highly conserved between species (Krieger et al, 2003), and an odorant-binding subunit. The numbers of odorant binding subunits vary widely between species. For example, head lice contain as few as 10 OR genes while ants contain as many as 300 OR genes (Kirkness et al, 2010; Smith et al, 2011).

OBPs are small (*ca.* 150 amino acids), water soluble proteins secreted into the lymph fluid surrounding the OSNs. The first OBPs were identified as pheromone binding proteins in moths (Vogt and Riddiford, 1981). They are widely believed to bind different odorants, solubilize and transport the odorants across the lymph fluid to the odorant receptors in the dendrites (Vogt and Riddiford, 1981; Pelosi and Maida, 1995). In addition to this major function, OBPs have been proposed to protect odorants from degradation or deactivation by odorant-degrading enzymes (Leal, 2013). A 2016 study by Larter et al. revealed a novel role for OBPs in buffering sudden changes in the odorant levels of the environment, suggesting that some sensilla may not require OBPs for odorant transport. OBPs are categorized into classes ("Classic", "C-minus", "C-plus" and "Duplex or Dimer") based on the number and pattern of cysteines. The six conserved cysteines of Classic OBPs form three disulphide bonds that stabilize their tertiary structure and help define a hydrophobic binding cavity (Hekmat-Scafe et al, 2002; Zhou et al, 2004; Vieira et al, 2007; Gong et al, 2009; Vieira and Rozas, 2011; Manoharan et al, 2013).

The evolution of insect chemosensory genes is characterized by an extremely rapid process of birth-and-death evolution, whereby genes are frequently added to the repertoire by gene duplication and neofunctionalization, and losses by pseudogenization and deletion (Sánchez-Gracia et al, 2009). Comparative genomic analysis showed that insect chemosensory genes appear in tandem arrays and likely originate by unequal crossing over, and evolve independently from each other (Vieira et al, 2007; Nozawa and Nei, 2007; Sánchez-Gracia et al, 2009; Eyun et al, 2017). The chromosomal clustering of OBPs was first observed in *Drosophila melanogaster* (Hekmat-Scafe et al, 2002; Vogt et al, 2002). Across the genomes of 12 *Drosophila* species, 69% of OBP genes are arranged in ten clusters (Vieira et al, 2007). In *B. mori*, the OBP genes are not randomly distributed in the genome, and most of them appear to be in close physical proximity within a specific region, for example, 20 OBPs are clustered on five chromosomes (Gong et al, 2009). The inter-chromosomal translocation events seem to be more frequent in ORs than OBPs, and ORs appear to be more scattered in the genome of *Drosophila* (Robertson et al, 2003). Different numbers of ORs in insect genomes can be explained by the differential odorants they need to detect; as an example, social insects such as ants may need more ORs for their social communications (Slone et al, 2017).

Heliothine moths (Lepidoptera: *Heliothinae*) include some of the world's most devastating pest species. *Chloridea virescens* (previously *Heliothis virescens*), also known as tobacco budworm, is an important agricultural pest in many crops in the United States, attacking cotton, flax, soybean and tobacco in its larval stages. The other two closely related Heliothine species, *Helicoverpa zea* and *Helicoverpa*

armigera, are the dominant agricultural pests in the Americas, and Europe, temperate Asia and Africa and Australia, respectively. Although *C. virescens*, *H. zea* and *H. armigera* are all polyphagous, with overlapping host plant use, *C. virescens* (tobacco budworm) shows a preference for laying eggs on Solanaceae (e.g., tobacco and tomato), while the other two *Helicoverpa* do not (Cunningham and Zalucki, 2014).

In order to study the evolutionary divergence of OR and OBP gene families among members of the Heliothinae, I manually curated OR and OBP genes in the publicly available reference genome of *C. virescens* (GenBank assembly accession: GCA_002382865.1; Fritz et al. 2018). The number, positions, and sequences of ORs and OBPs were identified. I then classified OBP genes into three different types (Classic, Minus-C and Plus-C), according to the number and position of conserved cysteine residues. In addition, phylogenetic relationships among chemosensory genes from the Heliothinae were examined to identify gene duplication, loss and sequence evolution and infer the gene family evolutionary histories.

3.2 Materials and Methods

3.2.1 Manual curation

The webapollo platform hosted by the insect 5000 genomes initiative (i5k; Poelchau et al, 2014) was used for manual curation of ORs and OBPs in the reference genome of *C. virescens* (GenBank assembly accession: GCA_002382865.1; Fritz et al. 2018). Publicly available OR and OBP proteins from closely related species, *H. zea*, *H. armigera* and *B. mori*, and some chemosensory receptor proteins of *C. virescens* were used to identify specific protein motifs or BLAST against the genome

assemblies. Existing gene models from a structural annotation were either accepted or modified in light of both these tblastn results and guidance from spliced RNA-seq reads from adult male and female antennae (GenBank accession numbers: SRR6417884 and SRR6417880). The genes were named according to the best matched ORs/OBPs of *H. zea* and *H. armigera,* which were identified and manually curated by Pearce et al (2017). The predicted *C. virescens* OBP protein sequences were aligned to one another with ClustalW (Thompson et al, 2003) and were classified according to the conserved number and spacing of cysteine residues in the peptide sequences (Gong et al, 2009; Vieira and Rozas, 2011; Manoharan et al, 2013).

3.2.2 Phylogenetic analysis

To study the evolutionary divergence of OR and OBP gene families among members of the Heliothinae, I compared the OR and OBP protein sequences from *C. virescens* to closely-related Heliothine species: *C. subflexa, H. zea, H. armigera* and *H. assulta.* In Pearce et al (2017), 83 ORs, one Orco and 40 OBPs, and 81 ORs, one Orco and 40 OBPs had been manually annotated in the reference genomes of *H. armigera* and *H. zea* (GenBank assembly accessions: GCA_002156985.1, GCA_002150865.1), respectively. For phylogenetic analysis, four and eight ORs of *H. armigera* and *H. zea,* respectively, were removed due to fragmentation of their protein sequences. In addition to the 80 annotated *C. virescens* ORs, the protein sequence of HvirOR11 (GenBank accession number: QLF97429.1), which could not be found in the assembly, was also included in the phylogenetic analysis. Six

pheromone receptors (OR6, OR11, OR13-16) and Orco from *C. subflexa* were added (Cao et al, 2021). The identified full-length ORs from *B. mori* (Tanaka et al, 2009; Liu et al, 2014) and *H. assulta* (Jiang et al, 2014; Xu et al, 2015; Wu et al, 2019) were also added.

Four fragmented OBP protein sequences were also removed from each of *H. zea* and *C. virescens* prior to phylogenetic analysis of the OBP gene family. The identified full-length OBPs from *B. mori* (Gong et al, 2009; Vieira and Rozas, 2011) and *H. assulta* (Li et al, 2009; Sun et al, 2012; Chang et al, 2017) were used for analysis. Because no pheromone binding proteins (PBPs) were annotated from the reference genomes of *H. armigera* and *H. zea,* publicly available protein sequences of *H. armigera* PBP 1-3 (GenBank accession numbers: CAC08212.1, ACD01993.1, AAO16091.1) and *H. zea* PBP (GenBank accession number: AAC36315.1) were identified and downloaded from GenBank for the phylogenetic analysis.

The protein sequences of the selected ORs and OBPs were aligned using MUSCLE algorithm (Edgar, 2004), and a maximum likelihood phylogenetic tree was constructed in MEGA 11 (Tamura et al, 2021) using Jones-Taylor-Thornton amino acid substitution model (JTT). To infer node support, 1,000 bootstrap replicates were calculated. The resultant tree was plotted using iTOL (https://itol.embl.de/; Letunic & Bork, 2021). Potential gene duplication and loss events were identified from the trees according to the number of corresponding orthologs, under the assumption that every species should have one ortholog for each gene.

3.3 Results and Discussion

3.3.1 Manual curation

In total, I manually annotated 80 ORs, one Orco and 49 OBPs from the *C. virescens* reference genome (GenBank assembly accession: GCA_002382865.1; Fritz et al, 2018) and named them according to the best matched ORs/OBPs of *H. zea* and *H. armigera* (Pearce et al, 2017). The OBP genes were clustered in the reference genome of *C. virescens*. The 49 OBPs were distributed across 15 scaffolds, and almost half of them (24 out of 49) were clustered within a 100 kb region of scaffold NWSH01000044.1 (Figure 3.1). The same clustering of OBP genes was also observed in *D. melanogaster* (Hekmat-Scafe et al, 2002; Vogt et al, 2002) and *B. mori* (Gong et al, 2009). OR genes appeared more scattered throughout the genome, having only a few clusters. An explanation for the multiple, widely dispersed gene clusters of *C. virescens* ORs is that inter-chromosomal translocation events are more frequent in ORs than OBPs (Conceição and Montserrat, 2008). An alternative hypothesis is that the distribution is related to gene regulation and expression, which was observed in vertebrate ORs (Sullivan et al, 1996; Young and Trask, 2002). For example, OR gene regulation might involve locus choice, gene choice within a cluster, and allele choice.

Chloridea virescens has more identified OBPs than other annotated Heliothine species: *H. zea* and *H. armigera* both have 40 OBPs in their reference genomes (Pearce et al, 2017). A model species in Lepidoptera, *B. mori*, has 44 identified OBPs and 43 of them are full-length (Gong et al, 2009; Vieira and Rozas, 2011). One possible reason for the increase in *C. virescens* is the different qualities of their

reference genomes. The reference genome of *C. virescens* is more fragmented than the other two *Helicoverpa* species, and some of the OBPs may not reflect true gene duplication events, but rather come from alternative haplotype-specific contigs. Future work could focus on testing the two hypotheses, for example through checking the read depth and nucleotide sequence divergence of the corresponding OBPs.

3.3.2 Classification of *C. virescens* OBP

Insect OBPs are divergent and the overall pairwise sequence similarity is modest (Hekmat-Scafe et al, 2002). The spacing pattern of the six conserved cysteines is the most typical feature of classical insect OBPs, and similar patterns have been found in *Drosophila* and *B. mori* (Hekmat-Scafe et al, 2002; Zhou et al, 2004; Vieira et al, 2007; Gong et al, 2009). The alignment of the full-length classical OBP proteins showed the pattern and positions of the six conserved cysteines (Figure 3.2). Following the naming proposed by Hekmat-Scafe et al (2002), we identified three types of OBPs in *C. virescens:* 34 Classic OBPs that have the six conserved cysteines, eight Minus-C and seven Plus-C OBPs that have fewer or more than six conserved cysteine residues, respectively (Table 3.2).

3.3.3 Phylogenetic analysis

The phylogenetic trees of ORs and OBPs were constructed using the maximum likelihood method with 1,000 bootstraps. In the phylogenetic tree of ORs, the Orco proteins are clustered as expected (Figure 3.3A, highlighted in light red). Orco is a special OR (also known as OR83b) that is highly conserved in most insect species, so the protein sequences from different species are very similar. In addition,

pheromone receptors from the Heliothine species are clustered but distantly related to *B. mori*'s ORs (Figure 3.3A, highlighted in light purple), which suggests that gene expansion and divergence occurred in the Heliothine species and *B. mori* separately. In the Heliothine moths, females create pheromone blends with some of the same volatile compounds used in different ratios. Males can respond to the pheromone compounds, but the strength of the response may vary, according to physiological or functional assays. Because these pheromone receptors can respond to some of the same ligands, there is also some sequence similarity at the protein level. Therefore, I expected to see pheromone receptor genes from different Heliothine species cluster in the phylogeny. Notably, in the pheromone receptor (PR) clade (Figure 3.3B), OR14 has two copies in *H. armigera*, *H. zea* and *H. assulta*, but only one copy in the *C. virescens* genome. There could be a potential gene loss in *C. virescens* and *C. subflexa*, or a gene duplication in the *Helicoverpa* species. A recent gene duplication of OR6 was identified in *C. virescens*, and the two copies are very similar (identity of protein sequences = 86.48%) and next to each other in the genome. However, considering the fragmented reference genome and poor structural annotation in the region, further studies will be needed to verify these findings.

The pheromone binding protein (PBP) sub-family of OBPs contains proteins that bind and solubilize the hydrophobic pheromone components and transfer them to the specific pheromone receptors. In the phylogenetic tree of OBPs, PBPs and general odorant binding proteins (GOBPs) placed in the same clade (Figure 3.4, highlighted in light red), which was also observed in *B. mori* (Gong et al, 2009). Surprisingly, no PBPs were annotated from *H. armigera* and *H. zea* by Pearce et al (2017), although

previous studies have identified several PBPs in *Helicoverpa* species using RT-PCR (Callahan et al, 2000; Wang et al, 2004; Zhang et al, 2012; Guo et al, 2012; Dong et al, 2017). Blast results demonstrated that these PBPs were still present in the genomes of *H. armigera* and *H. zea* (GenBank assembly accessions: GCA_002156985.1, GCA_002150865.1), but not annotated in the manual curation, therefore, we manually added HarmPBP 1-3 and HzeaPBP to the phylogenetic analysis. An explanation of the missing PBPs by Pearce et al (2017) is the quality limitation of the reference genomes of *H. armigera* and *H. zea*. With a fragmented short-read reference genome, these genes could be fragmented or missing from the identified genes.

In addition, we identified a gene expansion in the "Minus-C" group of *B. mori* (OBP 25-27) in the phylogenetic tree (Figure 3.4, highlighted in light blue). All of them are clustered in a small region of chromosome 5 in the genome of *B. mori* (Gong et al, 2009), indicating a recent rapid gene expansion and diversification. Another rapid gene expansion was observed in the Heliothine species, next to OBP11 of *B. mori* (Figure 3.4, highlighted in light blue). The OBPs from Heliothine species are more closely related than *B. mori's* OBP, which suggests a gene expansion and diversification in the Heliothinae moths. Among them, HvirOBP50 (a unique OBP in *C. virescens*) and HvirOBP4 are next to each other in the genome, indicating a recent gene duplication in *C. virescens*.

3.4 Conclusion

In summary, in this chapter I annotated 80 ORs, one Orco and 49 OBPs from the *C. virescens* reference genome. Three types of OBPs were classified according to the pattern of six conserved cysteines: 34 classic OBPs, eight Minus-C OBPs, and seven Plus-C OBPs. In addition, a phylogenetic analysis was done to identify potential gene duplication and gene loss of OR and OBP gene families among members of the Heliothinae. Two novel gene expansion events were observed in the "Minus-C" group of *B. mori* and the Heliothine moths. These gene diversification events could be related to the changes of olfactory-mediated behaviors between the Heliothine moths.

Tables: Chapter 3

Table 3.1. The OR gene family annotated from the reference genome of *C. virescens.*

HvirOR	HzeaOR	Hz_annotation_id	Ha_annotation_id	HarmOR	Scaffold	Start	End	Direction
OR1	HzeaOR1	HzOG200691	HaOG200691	HarmOR1	NWSH01000176.1	234098	238032	-
OR2					NWSH01002961.1	3629	12151	-
missing	HzeaOR3	HzOG200973	HaOG200973	HarmOR3				
OR4	HzeaOR4	HzOG200977	HaOG200977	HarmOR4	NWSH01000145.1	114618	118386	+
OR5	HzeaOR5	HzOG200938	HaOG200938	HarmOR5	NWSH01002259.1	75484	84104	-
OR6	HzeaOR6	HzOG200931	HaOG200931	HarmOR6	NWSH01000007.1	26260	30918	+
OR6b	HzeaOR6	HzOG200931	HaOG200931	HarmOR6	NWSH01000007.1	34794	53163	+
OR7	HzeaOR7	HzOG200856	HaOG200856	HarmOR7	NWSH01000040.1	33005	36729	+
OR8	HzeaOR8	HzOG200985	HaOG200985	HarmOR8	NWSH01006510.1	6253	13277	-
OR9	HzeaOR9	HzOG200747	HaOG200747	HarmOR9	NWSH01002968.1	417	2775	-
OR10	HzeaOR10	HzOG200920	HaOG200920	HarmOR10	NWSH01007896.1	37220	39263	+
missing	HzeaOR11	HzOG201020	HaOG201020	HarmOR11				
OR12	HzeaOR12	HzOG200855	HaOG200855	HarmOR12	NWSH01000040.1	21714	24836	+
OR13					NWSH01001866.1	45142	49299	-
missing	HzeaOR13-like	HzOG215628	HaOG215628	HarmOR13-like				
OR14	HzeaOR14	HzOG200928	HaOG200928	HarmOR14				
OR15	HzeaOR15	HzOG200929	HaOG200929	HarmOR15				

HvirOR	HzeaOR	Hz_annotation_id	Ha_annotation_id	HarmOR	Scaffold	Start	End	Direction
OR16	HzeaOR16	HzOG200930	HaOG200930	HarmOR16				
OR17	HzeaOR17	HzOG200971	HaOG200971	HarmOR17	NWSH01002272.1	11811	24656	+
OR18	HzeaOR18	HzOG200915	HaOG200915	HarmOR18	NWSH01003453.1	3621	6057	+
OR19	HzeaOR19	HzOG200978	HaOG200978	HarmOR19	NWSH01004225.1	12716	18929	-
OR20	HzeaOR20	HzOG200982	HaOG200982	HarmOR20	NWSH01000786.1	17346	24481	+
OR21	HzeaOR21	HzOG200944	HaOG200944	HarmOR21	NWSH01002810.1	13233	15497	+
OR22	HzeaOR22	HzOG200939	HaOG200939	HarmOR22	NWSH01002259.1	85092	90794	-
OR23	HzeaOR23	HzOG200904	HaOG200904	HarmOR23	NWSH01005468.1	36874	42304	-
OR24	HzeaOR24	HzOG200972	HaOG200972	HarmOR24	NWSH01002098.1	10225	11881	+
missing	HzeaOR25	HzOG200988	HaOG200988	HarmOR25				
OR26	HzeaOR26	HzOG200850	HaOG200850	HarmOR26	NWSH01001847.1	38534	41144	+
OR27	HzeaOR27	HzOG200861	HaOG200861	HarmOR27	NWSH01001630.1	31210	35701	+
OR28	HzeaOR28	HzOG200822	HaOG200822	HarmOR28	NWSH01002174.1	17780	20561	+
OR29	HzeaOR29	HzOG200984	HaOG200984	HarmOR29	NWSH01005276.1	1444	3487	+
OR30	HzeaOR30	HzOG200919	HaOG200919	HarmOR30	NWSH01007896.1	30882	33394	+
missing	HzeaOR31	HzOG200779	HaOG200779	HarmOR31				
OR32	HzeaOR32	HzOG201024	HaOG201024	HarmOR32	NWSH01003434.1	24718	33623	+
OR33	HzeaOR33	HzOG200821	HaOG200821	HarmOR33	NWSH01002174.1	4726	14297	+
missing	HzeaOR34	HzOG200793	HaOG200793	HarmOR34				
OR35	HzeaOR35	HzOG200880	HaOG200880	HarmOR35	NWSH01003963.1	15963	20481	+
OR36	HzeaOR36	HzOG200946	HaOG200946	HarmOR36	NWSH01002232.1	5412	9570	-

HvirOR	HzeaOR	Hz_annotation_id	Ha_annotation_id	HarmOR	Scaffold	Start	End	Direction
OR37	HzeaOR37	HzOG200979	HaOG200979	HarmOR37	NWSH01000780.1	173644	179639	-
OR38	HzeaOR38	HzOG200917	HaOG200917	HarmOR38	NWSH01001579.1	123876	125677	+
missing	HzeaOR39	HzOG200916	HaOG200916	HarmOR39				
OR40	HzeaOR40	HzOG200839	HaOG200839	HarmOR40	NWSH01004938.1	9882	18813	+
OR41	HzeaOR41	HzOG200827	HaOG200827	HarmOR41	NWSH01001711.1	36170	42596	+
OR41-like	HzeaOR41-like	HzOG212112	HaOG212112	HarmOR41-like	NWSH01003130.1	35532	38975	+
OR42	HzeaOR42	HzOG200780	HaOG200780	HarmOR42	NWSH01002573.1	44937	45844	+
missing	HzeaOR43	HzOG200797	HaOG200797	HarmOR43				
OR44	HzeaOR44	HzOG201018	HaOG201018	HarmOR44	NWSH01002690.1	20048	27274	+
OR45	HzeaOR45	HzOG201021	HaOG201021	HarmOR45	NWSH01000683.1	39929	61566	+
OR46	HzeaOR46	HzOG200864	HaOG200864	HarmOR46	NWSH01000975.1	18821	24356	+
missing	HzeaOR47	HzOG200772	HaOG200772	HarmOR47				
missing	HzeaOR48	HzOG200783	HaOG200783	HarmOR48				
missing	HzeaOR49	HzOG200943	HaOG200943	HarmOR49				
OR50	HzeaOR50	HzOG200941	HaOG200941	HarmOR50	NWSH01002259.1	94789	101740	+
missing	HzeaOR51	HzOG200927	HaOG200927	HarmOR51				
OR52	HzeaOR52	HzOG200820	HaOG200820	HarmOR52	NWSH01002174.1	11848	14297	+
OR53	HzeaOR53	HzOG200932	HaOG200932	HarmOR53	NWSH01006314.1	28982	38256	+
OR54	HzeaOR54	HzOG200914	HaOG200914	HarmOR54	NWSH01003519.1	43342	45897	-
OR55	HzeaOR55	HzOG200840	HaOG200840	HarmOR55	NWSH01001309.1	13478	14547	-
OR56	HzeaOR56	HzOG201026	HaOG201026	HarmOR56	NWSH01000442.1	154283	158788	-

HvirOR	HzeaOR	Hz_annotation_id	Ha_annotation_id	HarmOR	Scaffold	Start	End	Direction
OR57	HzeaOR57	HzOG201025	HaOG201025	HarmOR57	NWSH01003434.1	35521	41175	+
OR58	HzeaOR58	HzOG200947	HaOG200947	HarmOR58	NWSH01002695.1	10361	13284	-
OR59	HzeaOR59	HzOG201019	HaOG201019	HarmOR59	NWSH01002690.1	8461	11151	+
OR60	HzeaOR60	HzOG200799	HaOG200799	HarmOR60	NWSH01007359.1	10	2473	-
missing	HzeaOR61	HzOG200940	HaOG200940	HarmOR61				
OR62	HzeaOR62	HzOG200942	HaOG200942	HarmOR62	NWSH01002259.1	92113	93819	+
OR63	HzeaOR63	HzOG200798	HaOG200798	HarmOR63	NWSH01002961.1	256	1751	-
OR64	HzeaOR64	HzOG200918	HaOG200918	HarmOR64	NWSH01001579.1	109385	110552	+
missing	HzeaOR65	HzOG208187	HaOG208187	HarmOR65				
OR66	HzeaOR66	HzOG208188	HaOG208188	HarmOR66	NWSH01002174.1	97	4097	+
OR67	HzeaOR67	HzOG204092	HaOG204092	HarmOR67	NWSH01002968.1	4126	6253	-
OR68	HzeaOR68	HzOG201254	HaOG201254	HarmOR68	NWSH01000058.1	45666	50843	+
OR69	HzeaOR69	HzOG201234	HaOG201234	HarmOR69	NWSH01000302.1	58054	63219	+
OR70	HzeaOR70	HzOG214106	HaOG214106	HarmOR70	NWSH01001752.1	12007	26950	+
OR71	HzeaOR71	HzOG204587	HaOG204587	HarmOR71	NWSH01001446.1	1535	10573	+
OR72	HzeaOR72	HzOG204580	HaOG204580	HarmOR72	NWSH01001574.1	21579	27879	+
missing	HzeaOR73	HzOG206392	HaOG206392	HarmOR73				
missing	HzeaOR74	HzOG200987	HaOG200987	HarmOR74				
missing	HzeaOR75	HzOG200989	HaOG200989	HarmOR75				
missing	HzeaOR76	HzOG200990	HaOG200990	HarmOR76				
missing	HzeaOR77	HzOG201793	HaOG201793	HarmOR77				

HvirOR	HzeaOR	Hz_annotation_id	Ha_annotation_id	HarmOR	Scaffold	Start	End	Direction
OR78	HzeaOR78	HzOG206527	HaOG206527	HarmOR78	NWSH01006065.1	43849	46374	+
missing	HzeaOR79	HzOG207186	HaOG207186	HarmOR79				
OR80					NWSH01002690.1	12748	14293	+
OR81					NWSH01002634.1	245	2165	+
OR82					NWSH01000442.1	142581	147694	-
OR83					NWSH01000229.1	6023	23022	+
OR84					NWSH01003458.1	7387	10221	-
OR85					NWSH01001625.1	77355	83869	+
OR86					NWSH01002690.1	211	3206	+
OR87					NWSH01003458.1	14488	17423	-
OR88					NWSH01007896.1	30882	36270	+
OR89					NWSH01004177.1	10494	13504	-
OR90					NWSH01006537.1	11301	13303	-
OR91					NWSH01002810.1	4027	8937	+
OR92					NWSH01002810.1	1200	2602	+
OR93					NWSH01002810.1	6443	8937	+
OR94					NWSH01003581.1	16312	17839	+
OR95					NWSH01000145.1	290063	301383	-
OR96					NWSH01003458.1	280	2178	-
OR97					NWSH01005768.1	2280	5895	+

Table 3.2. The OBP gene family annotated from the reference genome of *C. virescens* and their classifications

HvirOBP	HzeaOBP	Hz_annotation_id	Ha_annotation_id	HarmOBP	Scaffold	Start	End	Direction	Length (aa)	Type
GOBP1	HzeaGOBP1	HzOG200773	HaOG200773	HarmGOBP1	NWSH01003619.1	7268	7915	-	166	Classic
GOBP2	HzeaGOBP2	HzOG200774	HaOG200774	HarmGOBP2	NWSH01000361.1	112098	114160	+	164	Classic
OBP1	HzeaOBP1	HzOG200804	HaOG200804	HarmOBP1	NWSH01000044.1	115144	117327	+	149	Classic
OBP2	HzeaOBP2	HzOG200790	HaOG200790	HarmOBP2	NWSH01000044.1	74180	75575	+	144	Classic
OBP3	HzeaOBP3	HzOG200802	HaOG200802	HarmOBP3	NWSH01000044.1	96594	97721	+	149	Classic
OBP4	HzeaOBP4	HzOG200806	HaOG200806	HarmOBP4	NWSH01000044.1	140536	141903	+	116	MinusC
OBP5	HzeaOBP5	HzOG200805	HaOG200805	HarmOBP5	NWSH01000044.1	130415	138664	+	149	Classic
OBP6	HzeaOBP6	HzOG200803	HaOG200803	HarmOBP6	NWSH01000044.1	100824	105689	+	149	Classic
OBP7	HzeaOBP7	HzOG200791	HaOG200791	HarmOBP7	NWSH01000044.1	79096	80676	+	150	Classic
OBP8	HzeaOBP8	HzOG200787	HaOG200787	HarmOBP8	NWSH01000044.1	61446	62694	-	123	Classic
OBP10	HzeaOBP10	HzOG200809	HaOG200809	HarmOBP10	NWSH01000044.1	157750	160797	+	138	MinusC
OBP12	HzeaOBP12	HzOG200644	HaOG200644	HarmOBP12	NWSH01004598.1	21615	23142	-	153	PlusC
OBP13	HzeaOBP13	HzOG200969	HaOG200969	HarmOBP13	NWSH01000208.1	182287	185013	+	143	Classic
OBP14	HzeaOBP14	HzOG200841	HaOG200841	HarmOBP14	NWSH01008731.1	3203	3811	+	121	MinusC
OBP15	HzeaOBP15	HzOG200853	HaOG200853	HarmOBP15	NWSH01001428.1	67007	68231	+	170	Classic
missing	HzeaOBP16	HzOG200643	HaOG200643	HarmOBP16						
OBP17	HzeaOBP17	HzOG200863	HaOG200863	HarmOBP17	NWSH01001889.1	3640	4149	+	139	Classic
OBP18	HzeaOBP18	HzOG200788	HaOG200788	HarmOBP18	NWSH01000044.1	67705	71824	+	148	Classic
OBP19	HzeaOBP19	HzOG200786	HaOG200786	HarmOBP19	NWSH01000044.1	52149	53682	-	151	Classic
OBP22	HzeaOBP22	HzOG200789	HaOG200789	HarmOBP22	NWSH01000044.1	72170	73555	+	148	Classic
OBP25	HzeaOBP25	HzOG200808	HaOG200808	HarmOBP25	NWSH01000044.1	154235	157424	+	149	Classic

HvirOBP	HzeaOBP	Hz_annotation_id	Ha_annotation_id	HarmOBP	Scaffold	Start	End	Direction	Length (aa)	Type
OBP26	HzeaOBP26	HzOG200807	HaOG200807	HarmOBP26	NWSH01000044.1	147121	153027	-	156	Classic
OBP27	HzeaOBP27	HzOG200800	HaOG200800	HarmOBP27	NWSH01000044.1	85990	88126	+	143	Classic
OBP28	HzeaOBP28	HzOG200801	HaOG200801	HarmOBP28	NWSH01000044.1	90124	92889	-	154	Classic
missing	HzeaOBP31	HzOG200796	HaOG200796	HarmOBP31						
OBP32	HzeaOBP32	HzOG200781	HaOG200781	HarmOBP32	NWSH01003491.1	13897	17386	+	228	PlusC
OBP33	HzeaOBP33	HzOG200967	HaOG200967	HarmOBP33	NWSH01000549.1	89636	91519	-	145	MinusC
OBP34	HzeaOBP34	HzOG200966	HaOG200966	HarmOBP34	NWSH01000549.1	93722	96167	-	137	MinusC
OBP35	HzeaOBP35	HzOG211753	HaOG211753	HarmOBP35	NWSH01000509.1	96400	97035	+	214	Classic
OBP36	HzeaOBP36	HzOG200881	HaOG200881	HarmOBP36	NWSH01000018.1	120612	122013	-	174	Classic
OBP37	HzeaOBP37	HzOG200925	HaOG200925	HarmOBP37	NWSH01006551.1	16547	22213	+	194	Classic
missing	HzeaOBP38	HzOG200818	HaOG200818	HarmOBP38						
OBP39	HzeaOBP39	HzOG201016	HaOG201016	HarmOBP39	NWSH01000610.1	40226	74140	-	187	Classic
missing	HzeaOBP41	HzOG200642	HaOG200642	HarmOBP41						
OBP42	HzeaOBP42	HzOG200636	HaOG200636	HarmOBP42	NWSH01006041.1	2006	4853	-	203	PlusC
OBP43	HzeaOBP43	HzOG200638	HaOG200638	HarmOBP43	NWSH01006041.1	10877	12402	-	139	Classic
missing	HzeaOBP44	HzOG200640	HaOG200640	HarmOBP44						
OBP45	HzeaOBP45	HzOG201656	HaOG201656	HarmOBP45	NWSH01002197.1	107188	108457	-	151	PlusC
missing	HzeaOBP46	HzOG201657	HaOG201657	HarmOBP46						
OBP47	HzeaOBP47	HzOG206450	HaOG206450	HarmOBP47	NWSH01000044.1	49758	50650	-	133	Classic
OBP48					NWSH01000044.1	83837	85196	+	150	Classic
OBP49					NWSH01000044.1	109391	110854	+	149	Classic
OBP50					NWSH01000044.1	137478	138664	+	90	MinusC
OBP51					NWSH01000044.1	145015	146916	+	152	Classic

HvirOBP	HzeaOBP	Hz_annotation_id	Ha_annotation_id	HarmOBP	Scaffold	Start	End	Direction	Length (aa)	Type
OBP52					NWSH01000361.1	130050	131132	-	167	Classic
OBP53					NWSH01000361.1	121965	123080	+	123	MinusC
OBP54					NWSH01000361.1	125279	126082	+	166	Classic
OBP55					NWSH01000361.1	119429	120561	+	172	Classic
OBP56					NWSH01000044.1	70164	71824	+	144	Classic
OBP57					NWSH01006551.1	2037	2509	+	105	MinusC
OBP58					NWSH01000044.1	111807	113195	+	149	Classic
OBP59					NWSH01000044.1	118033	120613	+	144	Classic
OBP60					NWSH01004598.1	5443	7092	-	164	PlusC
OBP61					NWSH01004598.1	1685	4806	-	179	PlusC
OBP62					NWSH01004598.1	11866	13459	-	180	PlusC

Figures: Chapter 3

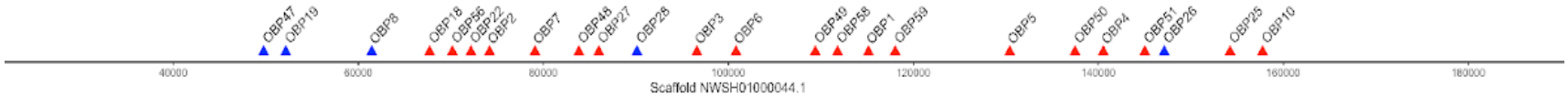


Figure 3.1. OBP gene cluster present in scaffold NWSH01000044.1 (scaffold length = 319,136 bp). Genes are depicted as triangles: red: forward transcription (+); blue: reverse transcription (-).

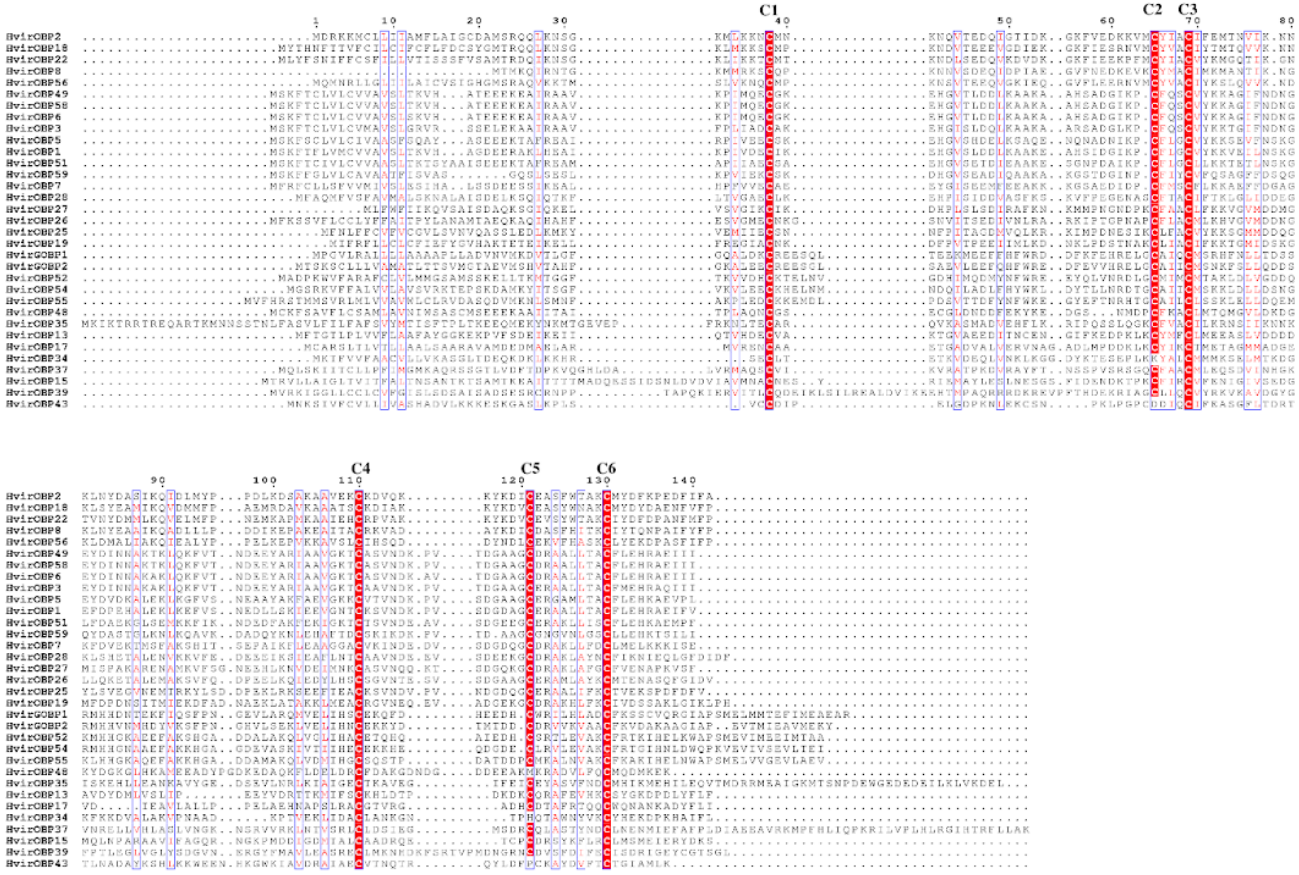


Figure 3.2. Alignment of the full-length classical OBPs in *C. virescens*. The protein sequences were aligned by ClustalW. Conserved residues are shown in red, and the six highly conserved cysteines are highlighted as red and marked as C1-C6.

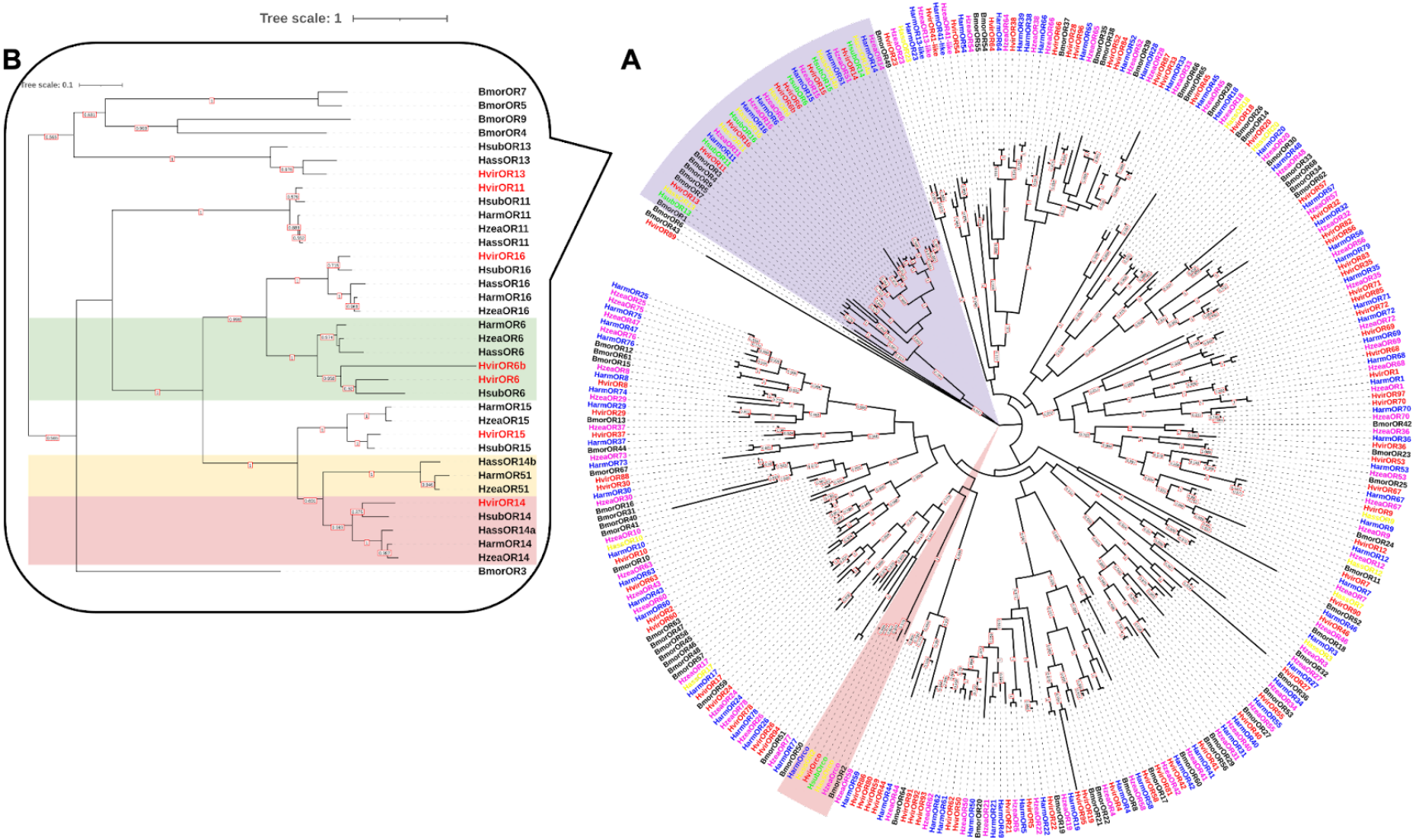

Tree scale: 1
B
Tree scale: 0.1
BmorOR7
BmorOR5
BmorOR9
BmorOR4
HsubOR13
HassOR13
HvirOR13
HvirOR11
HsubOR11
HarmOR11
HzeaOR11
HassOR11
HvirOR16
HsubOR16
HassOR16
HarmOR16
HzeaOR16
HarmOR6
HzeaOR6
HassOR6
HvirOR6b
HvirOR6
HsubOR6
HarmOR15
HzeaOR15
HvirOR15
HsubOR15
HassOR14b
HarmOR51
HzeaOR51
HvirOR14
HsubOR14
HassOR14a
HarmOR14
HzeaOR14
BmorOR3
A

Figure 3.3. Unrooted maximum likelihood phylogenetic tree of *C. virescens* ORs with other Lepidoptera ORs. Tree scales represent the number of amino acid substitutions per site. (A) Phylogenetic tree based on OR protein sequences from *C. virescens, C. subflexa, H. zea, H. armigera, H. assulta* and *B. mori.* The *C. virescens* (red), *C. subflexa* (green), *H. zea* (pink), *H. armigera* (blue), *H. assulta* (yellow) members are shown. The pheromone receptor (PR) and Orco clades were highlighted in light purple and light red, respectively. (B) Phylogenetic tree based on PR protein sequences from *C. virescens, C. subflexa, H. zea, H. armigera, H. assulta* and *B. mori.* Red: *C. virescens*. The clades of OR6, OR14a and OR14b were highlighted as light green, light red and light yellow, respectively. The bootstrap values of the branches were indicated on the nodes. Hvir: *C. virescens*, Hsub: *C. subflexa*, Hzea: *H. zea*, Harm: *H. armigera*, Hass: *H. assulta*, Bmor: *B. mori.*

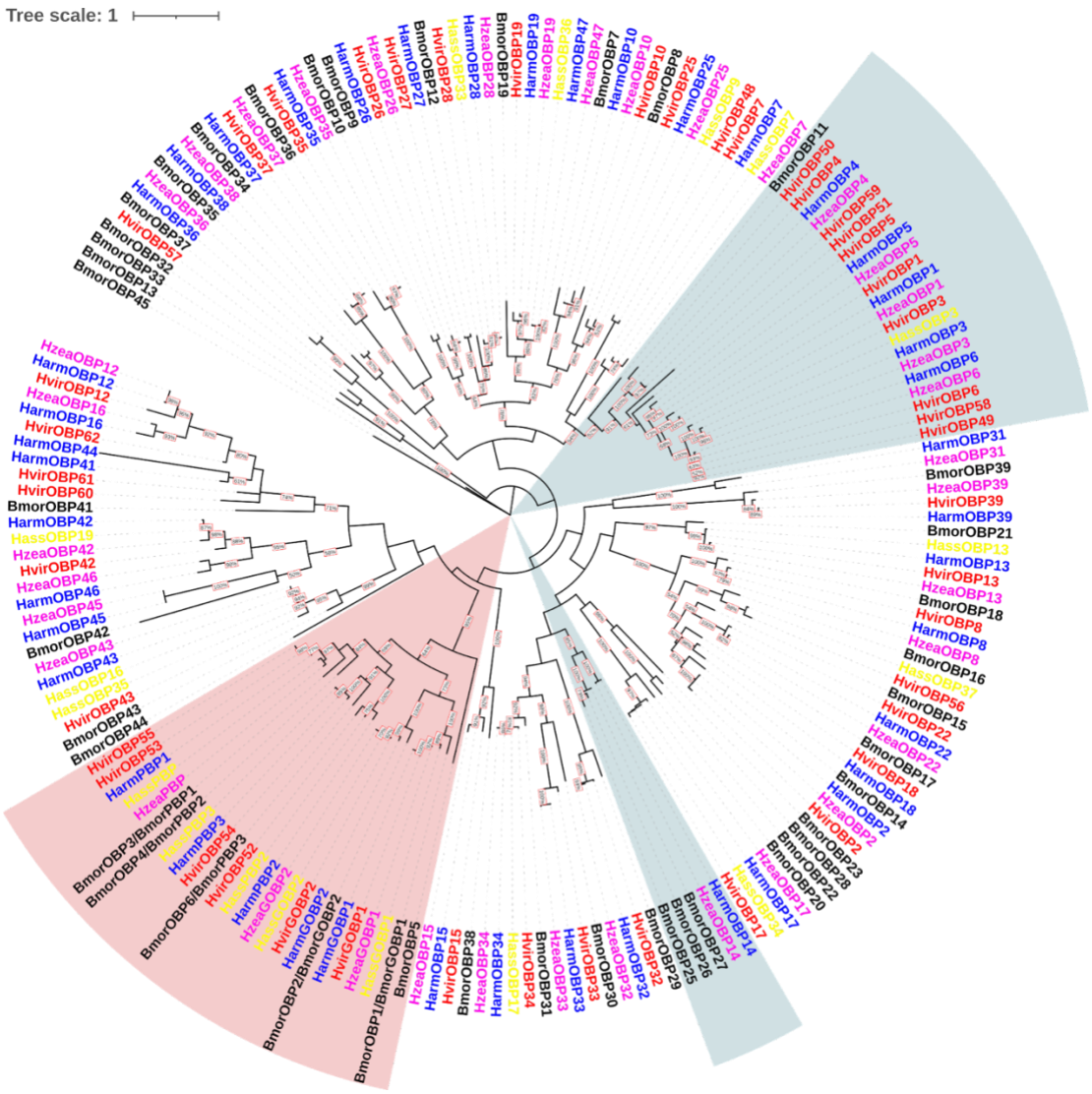


Figure 3.4. Unrooted maximum likelihood phylogenetic tree based on OBP protein sequences from *C. virescens* (red), *H. zea* (pink), *H. armigera* (blue), *H. assulta* (yellow) and *B. mori.* Tree scale represents the number of amino acid substitutions per site. The PBP/GOBP clade was highlighted in light red, and two gene diversification events were highlighted in light blue. The bootstrap values of the branches were indicated on the nodes. Hvir: *C. virescens*, Hzea: *H. zea*, Harm: *H. armigera*, Hass: *H. assulta*, Bmor: *B. mori.*

Chapter 4: Genetic variation at the pheromone receptors of *Chloridea virescens* and *Chloridea subflexa* and its association with host plant species

4.1 Introduction

The rules that govern sexual selection and speciation are complex and under some conditions they "contradict" one another, particularly for organisms that rely on pheromone signaling to identify mates. In Lepidoptera, sexual communication relies on a female to make a species-specific blend of volatiles known as pheromones that are detected by responding males. Males that do not respond to this species-specific blend are unlikely to find a mate, resulting in strong purifying selection (Karlson & Lu¨scher, 1959; Wyatt, 2003; Johansson & Jones, 2007). Furthermore, females whose pheromone blends are suboptimal should be much less likely to attract a responding male (Phelan 1992). This system reinforces species boundaries, and yet somehow speciation events have occurred in the Lepidoptera. The mechanisms by which male and female moths can circumvent this putatively strong purifying selection and diverge to become new species is unclear. One possibility is that host plant preference plays a role in speciation of Lepidoptera, and this has been suggested for other insect groups (Light et al, 1993; Feder et al, 1998; Yang et al, 2004; Deng et al, 2004). It has been known for some time that host plant odors synergize with pheromone blends in Lepidoptera species to increase male attraction and offspring success (Ochieng et al, 2002; Yang et al, 2004). Furthermore, although pheromone receptors (PRs) are

thought to function to optimally detect female pheromones (Krieger et al, 2004), two recent studies showed that they can also elicit electrophysiological responses to host plant odors (Lebreton et al, 2017; Rouyar et al, 2015). Perhaps selection for shifts in host plant use result in polymorphisms in pheromone perceiving ORs, which ultimately leads to speciation.

Two closely related species, *Chloridea (Heliothis) virescens* and *Chloridea (Heliothis) subflexa*, have different preferred host plants and different optimal pheromone blends. *C. virescens* is a polyphagous crop pest, attacking such crops as tobacco, cotton, flax, soybean, etc. While *C. subflexa* is a specialist, only feeding on a small number of plants in the *Physalis* genus. Yet they can be hybridized in the laboratory, and males weakly respond to the pheromone blend of the opposite species (Groot et al, 2006). The PRs and their expression levels are known (Vasquez et al, 2011). Vasquez et al (2011) tested the expression levels of OR6, OR14-16 in both *C. virescens* and *C. subflexa* by qRT-PCR and found significant interspecific differences in expression level of OR14-16. OR6 was expressed at comparable levels between species, however. Significant male-biased intraspecific expression was observed in OR6, OR14 and OR15 from both *C. virescens* and *C. subflexa*, and in contrast, OR16 was expressed at similar levels in males and females. A recent study by Cao et al (2021) functionally characterized the PRs in *C. subflexa*, and they found that a single mutation in OR6 can alter its function of *C. virescens*, transitioning from a response to both Z9-16:Ald and Z9-14:Ald to one to only Z9-16:Ald, a similar function as that of *C. subflexa*'s OR6. However, the reverse mutation could not change the function of *C. subflexa* to that of *C. virescens*. Because the discovery of this mutation occurred

using long-term laboratory reared populations, screening wild populations for similar mutations is of interest. Such screening may reveal novel nucleotide sequence variation present in these receptor genes under natural conditions.

Host plant odors synergize with species-specific pheromones to enhance male attraction, in a number of lepidopteran species (e.g., *Helicoverpa zea*, Ochieng et al, 2002; *Cydia pomonella*, Yang et al, 2004, Ansebo et al, 2005; *Spodoptera litura*, Fang et al, 2018), yet it is unclear whether this is the case for *Chloridea* species. Lepidopteran females generally use plant volatiles to identify suitable plant hosts for oviposition, and often have preference for some hosts over others. These preferences, which likely have a genetic basis (Oppenheim et al, 2012; Oppenheim et al, 2018), will be inherited by offspring. Moreover, a preliminary study by Dr. M.L. Fritz suggested that *C. virescens* from Maryland prefer to oviposit on tobacco (2/3 plants infested with larvae), followed by chickpeas (1/3 plants infested) and cotton (No plants infested). There is evidence that *C. virescens* can also use *Physalis* species, even tomatillo, as a plant host (Sitchawat and Thurston 1980, Laster et al, 1982, Oppenheim et al, 2018), but this and other species in the genus *Physalis* are the preferred hosts of *C. subflexa* (Oppenheim et al, 2012). *C. virescens* is not known to feed on *Physalis* in nature, and the survival rate of *C. virescens* on *Physalis angulata* is only 10% of *C. subflexa*'s in laboratory assays (Oppenheim et al, 2012).

To examine whether the genetic variation of PRs is associated with host plant choice, we evaluated the intra-specific genetic diversity of PRs in the field collected populations of *C. virescens* and *C. subflexa* and tested its association with their host plant choices. To understand when and where we can collect these *Chloridea* species

in Maryland, we initially quantified host plant choice by comparing larval densities and infestation rates of *C. virescens* and *C. subflexa* in tobacco, tomatillo, chickpeas and non-Bt cotton at three research farms in Maryland. Our hypothesis was that *C. virescens* larval density would be highest on its preferred host plant, tobacco, followed by chickpeas, non-Bt cotton and tomatillo. We then quantified genetic variation in the PRs of wild *Chloridea* males trapped in preferred and non-preferred host plant fields and examined how the amino acid variants were distributed within each receptor. The strength and type of selection imposed on the pheromone receptor genes of each species was inferred from our sequencing data, and we found that OR6 is under strong species-specific purifying selection. Using AMOVA, we tested the hypothesis that some standing genetic variations in PRs may be associated with host plant choices. The results indicated that in *C. virescens*, host plant associated population differentiation only existed in genes OR6, OR55, OR66 and OR78. In OR6, which serves as a PR that creates prezygotic isolation between *C. virescens* and *C. subflexa*, the most highly divergent sites were all in introns.

4.2 Materials and methods

4.2.1 Agricultural resources

Plots of tobacco, tomatillo, chickpeas and non-Bt cotton were planted at each of three research farms in central Maryland: CMREC Beltsville (38° 59' 15.396" N, 76° 49' 13.8972" W) Upper Marlboro (38° 51' 35.9994" N, 76° 46' 48" W), and WMREC Keedysville (39° 30' 0" N, 77° 43' 11.9994" W). Plot sizes ranged from two to eight 100 ft rows with conventional row spacing (2-5 feet depending upon the

crop). Plot sizes varied according to crop, and plot orientation was randomized by farm and year. Non-Bt cotton and tomatillo were planted over sheets of black plastic, two rows in each plot with 2 feet between plants. Tobacco and chickpeas were planted in the bare ground, six rows and eight rows respectively, with 40 inches between rows and 2 feet between plants. Moth field collections were performed on a weekly basis in each farm and for eight weeks, beginning in the last week of June in 2020 and 2021.

Variation in germination rate is common in agricultural studies and can occur due to quality of germplasm or environmental conditions (e.g., drought, flooding). Therefore, we quantified and reported mid-season plot density using stand counts made in the 4th or 5th week of the experiment. For each plot, we used a random number generator to generate both a row number and distance from the edge of the plot. The number of plants within 5 ft of the distance in a row was counted, and this was repeated five times. For example, if the random number was 57.5, we would count the plants found in one row between 52.5 ft to 62.5 ft from the edge of the field. The total number of plants in each plot was estimated as the average of five samples/10 ft * the plot length in ft. Chi-squared tests were performed to test whether there were statistically significant differences in the plant densities between years and farms for each crop.

4.2.2 Random sampling of *Chloridea* larvae

We first studied host plant use by *C. virescens* and *C. subflexa* in Maryland. The goal of *Chloridea* larvae sampling was to compare their densities and quantify the infestation rates of different host plants. The plots were small enough to enable

completely random sampling, which provided the highest statistical independence among samples. Without data on the spatial distribution of the larvae, completely random sampling is most accurately used for density estimation. We examined 25 randomly sampled plants per crop per farm on a weekly basis. Because some plants were more difficult to sample than others due to their canopy shape, the amount of time examining all 25 plants in each plot was recorded on each sampling date.

A random number generator was used to generate 25 number pairs (a, b) for each crop on each sampling date, where a is a random row number and b is a random number between 1 and 100. The random number b was used as a distance to walk from the start of row a, and then the nearest plant was chosen. Different strategies were applied to sampling different host plants. In tobacco and chickpeas, we walked to each plant indicated by the random distance generator, searched the whole plant and used visual observations to collect all the *Chloridea* larvae. In non-Bt cotton and tomatillo, we first checked up to 20 fruits and then spent 60 seconds checking as many leaves as we could. Even though larvae were known to live in the fruits, leaves were also checked in case there were larvae moving in the leaves (e.g. foraging for fruits). Low numbers of *C. virescens* collected from non-Bt cotton in previous field studies, motivated us to visually screen every cotton plant in an attempt to collect as many *Chloridea* larvae as possible in the first 4 weeks. The number of *Chloridea* larvae on each plant for each crop was recorded, and all the larvae were collected and returned to the lab for rearing.

Infestation rate (IR) was calculated as the proportion of sampled plants that were infested with *Chloridea* larvae (IR = the number of plants that were infested/25

*100). The amount of time spent sampling each crop was also recorded for each collection day. Sometimes one plant was infested with more than one larva, and thus the number of infested plants was different from the total number of larvae collected from the plot. Therefore, we also calculated infestation rate according to units of sampling time (per minute), which accounted for the fact that some plants contained more than one larva.

Field-collected larvae were brought to the lab and reared in 1-2 oz. cups of meridic diet (Southland Products Inc., Lake Village, AR). Pupae were then transferred into 64 oz. circular plastic containers lined with white paper towel for emergence. Following emergence, adults were transferred to new containers and held in groups, and reared for three days with sugar water. Three day old adults were sacrificed 2-4 hours after the onset of scotophase and stored at -80°C prior to genetic studies. Flash-freezing of these adults was carefully synchronized to mid-schotophase, which represents the period of sexual activity for both *C. subflexa* and *C. virescens*. While of minor importance for the present study of nucleotide sequence variation between species, the timing of flash-freezing will be important for future gene expression studies of these field-collected samples.

4.2.3 Trap collection

The goal of the 8-week trap collection was to compare genetic variation in the PRs of wild *Chloridea* males trapped in different host plant fields. Previous studies evaluated the efficacies of pheromone-baited moth traps in capturing male *Helicoverpa* in Florida, and Scentry® *Heliothis* trap had modest performance (Guerrero et al, 2014). Scentry® *Heliothis* traps also captured significantly more

males of other Noctuidae moths than bucket traps in Mexico (Malo et al, 2001). The efficacy of a pheromone trap also depends on the type of pheromone lure(s). Previous evaluation of commercial pheromone lures revealed that Trécé and Scentry lures were equally effective in catching *Spodoptera exigua*, and the longevity of Scentry lures was larger than Trécé and Hercon lures (Lopez Jr, 1998). Other studies have shown that the type of substrates used for releasing the pheromone can influence trap captures, for example, more *Plutella xylostella* males were captured with pheromone mixtures released from gray septa than from red septa (Mayer and Mitchell, 1999). Considering their commercially availability and ease of handling, we used Scentry® 31'' Heliothis traps and Scentry® *C. virescens* (tobacco budworm) lures (red rubber septa) in our *Chloridea* trap collection.

Scentry® 31'' *Heliothis* traps and Scentry® *C. virescens* (tobacco budworm) lures (red rubber septa) were purchased online from Great Lakes IPM™ (https://www.greatlakesipm.com/). Scentry® *Heliothis* traps with a removable top chamber were installed at each farm and were mounted *ca*. 1-1.5 ft. above a tobacco or tomatillo plant 25 ft into the plot (Figure 4.1). Scentry® *C. virescens* lures were secured at the bottom of each trap. A control trap to assess trap catch in the absence of host plants was set *ca.* 1-1.5 ft. above mixed wild and cultivated grass species in each farm in 2021. Trap heights were adjusted as plants grew to maintain the distance of trap openings from plant foliage. The lures were exchanged every 2 weeks during the 8-week field collection period. Traps were checked twice per week, and the numbers of adults collected in each trap were counted. The adults were returned to the lab and stored in -4℃ for preservation and identification.

4.2.4 Species identification and whole genome sequencing (WGS)

Species were identified morphologically and with molecular tools. *Helicoverpa zea* can be separated from *Chlorideas* morphologically. Yet *Chloridea* moth specimens from traps can sometimes be dead, bleached and brittle, making them difficult to morphologically identify. Thus, *C. virescens* and *C. subflexa* were identified according to known nucleotide differences in the nuclear gene, elongation factor -1 (EF-1): *C. virescens* and *C. subflexa* can be distinguished from each other by 5 interspecific nucleotide variants (Appendix Figure 4.1). For each morphologically identified *Chloridea,* genomic DNA was isolated using a Qiagen DNeasy kit (Qiagen, Valencia, CA, USA). Novel primers targeting a region of the EF-1 gene were designed and used to perform PCR: forward primer: 5'-AGG GTA AGG CTG AAG GTA AAT G-3', reverse primer: 5'-CAC CAG ACT TGA TGG ACT TAG G-3'. Each 20-µL PCR mix was made with: dH_2O: 12.5 µL, dNTPs: 0.4 µL with a final concentration of 200 µM, forward primer: 1 µL at 10 µM, reverse primer: 1 µL at 10 µM, 5× GoTaq Reaction Buffer: 4 µL, GoTaq DNA Polymerase (Progema): 1 µL at 25 units/mL and 1µL of genomic DNA. A negative control was made with 1µL of dH_2O. PCR was performed under the following conditions: 95°C for 3 min, 34 cycles of (95°C for 30 s, 60°C for 30 s, 72°C for 2 min), 72°C for 5 min, 4°C hold. The PCR products were purified using exonuclease I and antarctic phosphate. Each 20-µL Exo-AP master mix was made with 1 µL exonuclease I, 1 µL antarctic phosphate, 1 µL antarctic phosphate buffer and 17 µL dH_2O. 6 µL PCR products were mixed with 2 µL Exo-AP master mix under the following conditions: 37°C for 15 min, 80°C for 15

min, 15°C hold. The purified PCR products were sent to GENEWIZ (https://www.genewiz.com/) for Sanger sequencing.

Ten *C. virescens* from both tobacco and tomatillo traps (20 total) were chosen to test the hypothesis that host plant choice can impact the population structure of *C. virescens.* Although we used *C. virescens* pheromone lures to bait our Scentry traps, *C. subflexa* males were also collected in our traps throughout the course of our study (see below). This presented an opportunity to test the hypothesis that *C. subflexa* male responders to *C. virescens* pheromone blends show unique nucleotide sequence variants at their PR genes relative to those from the broader population of wild *C. subflexa* that could be collected directly as larvae from tomatillo. Therefore, we compared ten wild *C. subflexa* adults collected from tomatillo plots as larvae and reared to adults in the lab (see section 3.2.2) with ten *C. subflexa* collected as adults from pheromone-baited traps to test the hypothesis that genetic variation in the PR genes of *C. subflexa* was associated with collection method (natural host *vs. C. virescens* lures). DNA extractions of 20 *C. virescens* and 20 *C. subflexa* (for collection information see Appendices Table 4.4) were sent to Maryland Genomics (Institute for Genome Sciences, University of Maryland School of Medicine) for Illumina library preparation and NovaSeq6000 next-generation sequencing with average sequencing depth of ~15X. Libraries were prepared using Kappa Hyper prep kits.

4.2.5 Structural annotation of *C. subflexa* genome assembly

The reference genome of *C. subflexa* (GenBank assembly accession: GCA_022398575.1) was generated in Chapter 2 of this thesis, but the structural and

functional annotation conducted here allowed for identification of the chemosensory receptors, including the PR genes. Prior to automated annotation, the genome was masked using RepeatMasker and RepeatModeler (Smit et al, 2013). A custom repeat library was constructed by RepeatModeler 2.0.2, and the repeats were filtered using transposonPSI (https://github.com/NBISweden/TransposonPSI) and BLASTx to remove repeats that were derived from the protein coding genes. Then, the repeat sequences were searched by RepeatMasker 4.0.7 with default parameters.

The protein coding genes were annotated by integrating the evidence of ab initio, transcriptome-based prediction and homology-based annotations using the Maker 3.01.04 pipeline (Cantarel et al, 2008), run iteratively. Firstly, RNA-seq data of *C. subflexa* from different tissues including gut, thorax, male antennae and female antennae were obtained from NCBI (SRA accession numbers ERR738599, ERR738600, ERR738601, ERR738602, SRR6417881, SRR6417879, SRR6417883). For each dataset, the RNA-seq data were assembled into cDNAs with Trinity 2.8.3 (Grabherr et al, 2011) or SPAdes 3.15.3 (Bushmanova et al, 2019), and all the assembled transcriptomes were compiled into one file. Secondly, to prepare for the homology-based evidence, protein sequences or the annotated gff files from *C. subflexa, C. virescens, H. zea* and *H. armigera* were downloaded from NCBI. In the first run, the protein coding gene structures were predicted by Maker 3.01.04 with the above evidence. Augustus 3.4.0 (Stanke et al, 2008) and SNAP (Korf, 2004) were automatically trained with the Maker output. In the second run, the marker derived gff3 file, SNAP created hmm file and Augustus gene prediction models from the first round were combined with the transcriptome- and homology-based evidence as input

files. A quantitative assessment of annotation completeness was conducted using BUSCO v4.0.6 (insecta_odb10; Seppey et al, 2019).

4.2.6 Identification of pheromone receptors

Pheromone receptors OR6, OR13, OR14 and OR16 were identified in the reference genomes of *C. virescens* (GenBank assembly accession: GCA_002382865.1; Fritz et al, 2018) and *C. subflexa* (GenBank assembly accession: GCA_022398575.1; Guo et al, 2022). Their mRNAs (Cao et al, 2021; GenBank accession numbers: MN399799.1, MN399801.1, MN399802.1, MN399804.1 of *C. virescens* and MN399806.1, MN399808.1, MN399809.1, MN399811.1 of *C. subflexa*) were obtained from NCBI to blast against the genome assemblies. No gaps were found in the ORs of *C. virescens* (as described in Chapter 2). In *C. subflexa*, gaps within each OR were filled with the mRNA sequences, when possible. *C. subflexa* OR6 was split across two scaffolds (scaffold_4031 and scaffold_489), but these were joined manually to become Scaffold_4031_rc.Scaffold_489, used in this work (Appendix Supplemental Data S4).

4.2.7 Examination of nucleotide diversity within ORs

The quality of the Illumina reads from each *C. virescens* and *C. subflexa* individual was checked by FastQC (Babraham Bioinformatics Cambridge, UK), and adapters were trimmed using Trimmomatic (version 0.33; Bolger et al, 2014) with parameters: simple clip threshold=7, seed mismatches=2, palindrome threshold=40, minimum sequence length=30, and training quality=20. Reads were mapped to the reference genomes of *C. virescens* (GenBank assembly accession:

GCA_002382865.1; Fritz et al, 2018) or *C. subflexa* (adapted from GenBank assembly accession: GCA_022398575.1; Guo et al, 2022) using Bowtie2 (version 2.3.4.1; Langmead & Salzberg, 2012). Prior to calling variants, we checked individual read count, mapping rate, GC content and mean read alignment depth across populations to ensure uniformity (Appendix Table 4.1). Two individuals (Hvtob2 and Hvtom3) in the *C. virescens* populations had lower than expected mapping rates (Hvtob2: 10.45%, Hvtom3: 28.08%) and were removed from the data set. Samtools (version 1.12; Li et al, 2009) view was used to convert SAM files output to BAM files, and variants were called by bcftools 1.7 (Danecek et al, 2021) in a Variant Call Formatted (VCF) file in each species. We used vcftools (version 0.1.15; Danecek et al, 2011) to filter the called variants prior to downstream population genomic analysis. The filtered dataset included loci that: 1) were sequenced to a depth of 3 or more reads, 2) had a minor allele frequency of 0.1 or greater, 3) were represented in at least 50% of individuals.

A common measure of genetic diversity, $\boldsymbol{\pi}$, was calculated as the proportion of nucleotides that differ per two randomly chosen DNA sequences, averaged across all pairwise comparisons (Nei and Li, 1979). Site $\boldsymbol{\pi}$ values were calculated by vcftools for PRs OR6, OR13, OR14, OR16 and a highly conserved co-receptor, Orco, in *C. virescens* and *C. subflexa* populations. Orco is highly conserved across most insect species, and we expected it to have low genetic diversity relative to other ORs. A custom R script was used to calculate $\boldsymbol{\pi}$ values of the whole genes, as well as the coding regions.

4.2.8 Comparison of amino acid variants and their distributions across *Chloridea* ORs

The coding sequences of OR6, OR13, OR14, OR16 and Orco were generated for each individual according to the called SNPs and were translated into proteins. In each species the proteins were aligned with ClustalW (https://www.genome.jp/tools-bin/clustalw; Thompson et al, 2003), and the variant sites were counted and recorded. Biostrings package of R (Pages et al, 2013) was used to count and compare the allele frequencies of nonsynonymous SNPs in *C. virescens* trapped from tobacco and tomatillo, *C. subflexa* trapped from tomatillo and wild *C. subflexa* collected from tomatillo plots. A Fisher's exact test was used to determine whether variations between host plants and/or collection approach were significant or not.

To examine how the amino acid sequence variation was distributed within the receptor proteins, transmembrane topology was predicted and plotted for each OR. The transmembrane domains (TMDs) were predicted using TMHMM server version 2.0 (http://www.cbs.dtu.dk/services/TMHMM/; Krogh et al, 2001) based on the posterior probabilities of inside/outside/TM helix for each residue. The topology diagrams were made by Protter (http://wlab.ethz.ch/protter/#; Omastis et al, 2014). The variant sites of each OR were highlighted as yellow. We initially calculated a ratio to quantify whether the distribution of amino acid variants was biased toward any one of the three specific regions of each receptor: the intracellular loops (IL), the transmembrane domains (TMD) and the extracellular loops (EL). This ratio was calculated as the number of observed amino acid variations per region divided the number of expected amino acid variants. Expected amino acid variants: total number of amino acid variants * length of the domain / length of the protein. A ratio of 1 indicates that amino acid changes occurred at the same rate across the entire protein.

Chi-square tests were based on the number of amino acid variants and used to determine statistical significance.

4.2.9 Analysis of molecular variance (AMOVA)

An analysis of molecular variance (AMOVA) was performed to quantify the distribution of genetic diversity between and within populations (Excoffier et al, 1992). AMOVA is similar to other approaches based on analyses of variance of gene frequencies (Weir and Cockerham, 1984), but it takes into account the number of mutations between molecular haplotypes. To determine whether host plants appear to be associated with genetic structure of *C. virescens*, and whether *C. subflexa* males attracted by *C. virescens* lures were different from wild *C. subflexa*, we carried out an AMOVA using package poppr (v.2.9.3; Kamvar et al., 2014) in R version 4.0.3 and on the VCF file of SNPs in OR genes. The cutoff was set to 0.3, which means loci containing missing values greater than 30% were ignored. A Monte Carlo test with 1,000 random permutations was used to determine the statistical significance of each factor in the AMOVA.

F_{st} values (Weir and Cockerham, 1984) were calculated to estimate genetic divergence for OR6 between the *C. virescens* populations. A Chi-squared test was made to test if the top 5% most highly divergent sites showed a biased distribution within the gene (e.g. if they were clustered in the introns). To further verify if these sites were linked, linkage disequilibrium (LD) values for these most highly divergent sites were calculated using Plink v1.9 (Purcell et al, 2007), and a heatmap was plotted by ggplot package (Wickham, 2016) in R.

4.2.10 Selection on PRs in wild *C. virescens* and *C. subflexa.*

Tajima's neutrality test was performed using the PopGenome package of R (Pfeifer et al, 2014) to determine whether the pattern of mutations of each OR followed assumptions of neutrality or was caused by non-random processes (e.g. selection or population demographic change). The VCF file of each OR was used as input. To demonstrate that the Tajima's D values are significantly different from neutrality, coalescent-based simulations were conducted using the coala package of R (Staab and Metzler, 2016) under neutral conditions. We conducted 1,000 simulations using a population demographic model that assumed diploid individuals with gene length of 5,000 bp, a mutation rate of $2.9x10^{-9}$ (Keightley et al, 2015; Taylor et al, 2021) and theta = $4*N_0*mu*5000$. Present day population sizes (N_0) varied as follows: 5,000, 10,000, 15,000, 20,000 (estimated from Blanco et al, 2007). Except for balancing selection, sudden population contraction could be another explanation for positive Tajima's D, so we also simulated a population whose size decreased from $N_0 = 20{,}000$ to 30% at half-the coalescent time (t = 0.5). Our p-value was calculated as the probability of observing a Tajima's D value greater than or equal to our empirically-derived Tajima's D value of 1.45 for OR6 in 1,000 simulations.

The maximum likelihood probabilities of predicted haplotypes in the field populations were estimated for OR6, OR13, OR14, OR16 and Orco using haplo.em of R package haplo.stats (Sinnwell et al, 2022). All haplotypes present in field populations were compared to those identified in the laboratory samples used by Cao et al. (2021). Since most of the haplotypes in Cao et al. (2021) cannot be found in the predicted haplotypes, the haplotype with the highest probability in our field

population was used for further phylogenetic analysis. Protein sequences of ORs from *C. virescens, C. subflexa, H. armigera, H. zea* and *H. assulta* were aligned using MUSCLE algorithm (Edgar, 2004), and a maximum likelihood phylogenetic tree was constructed in MEGA 11 (Tamura et al, 2021) using Jones-Taylor-Thornton amino acid substitution model (JTT). To infer confidence, 1,000 bootstrap replicates were calculated. The resultant tree was plotted using iTOL (https://itol.embl.de/; Letunic & Bork, 2021).

Considering functional divergence among gene paralogs or species orthologues, tests of selection were performed for the five OR orthologs separately. The ratio (ω) of the number of nonsynonymous substitutions per nonsynonymous site (dN) to the number of synonymous substitutions per synonymous site (dS) was calculated for each lineage using the codeml module of PAML 4.9 (Yang 2007). ω is expected to be > 1 only if natural selection is associated with changes in the protein sequence, indicating positive or adaptive selection; $\omega = 1$ indicates neutral selection and $\omega < 1$ indicates negative or purifying selection (Yang et al, 2000). The protein sequences of each OR ortholog were aligned and their phylogenetic relationship was predicted using ClustalW (Thompson et al, 1994), and pal2nal.pl (Suyama et al, 2006) was used to get a codon-based nucleotide alignment. ω values were calculated using the codeml module with Site model Model 0: one-ratio.

4.3 Results and discussion

4.3.1 Host plant use of *C. virescens* and *C. subflexa* in central Maryland

No significant differences in plant density among years and farms were found in tobacco, tomatillo and cotton at an α-level of 0.05 (tobacco: $\chi^2 = 2.90$, p-value = 0.23, tomatillo: $\chi^2 = 0.02$, p-value = 0.99, cotton: $\chi^2 = 3.07$, p-value = 0.22). Significant differences were observed for chickpeas, however ($\chi^2 = 223.58$, p-value < 0.001). This can be attributed to low chickpea germination and survival rates in Keedysville in 2021.

In total, 310 and 518 *Chloridea* larvae were collected in 2020 and 2021, respectively (Appendix Table 4.2 and 4.3). The overall trend for average infestation rates in 2020 and 2021 showed increases during the 8-week collection period (Figure 4.2). In 2020, about 52.21% *C. virescens* larvae were collected from tobacco, followed by chickpeas (46.02%) and cotton (1.77%). The average infestation rates of *C. virescens* larvae in tobacco, chickpeas and cotton were 7.50%, 7.33% and 0.33% with the order tobacco > chickpeas > cotton (Figure 4.3a). However, most *C. virescens* larvae (78.67%) were collected from chickpeas, followed by tobacco (19.47%) and cotton (1.87%) in 2021. Of them 72.88% were collected from Upper Marlboro, and the remaining 27.12% were all collected from Beltsville. No *C. virescens* were collected from chickpeas in Keedysville in 2021, likely due to low chickpeas germination and survival rate at that farm (see above). In 2021, the average infestation rates of *C. virescens* larvae in tobacco, chickpeas and cotton were 10.00%, 24.17% and 1.17% with the order chickpeas > tobacco > cotton (Figure 4.3b). The average infestation rates of *C. subflexa* in tomatillo were 20.33% and 15.33% in 2020

and 2021, respectively. Most of the *C. subflexa* larvae (88 of 122 in 2020, 109 of 143 in 2021) were collected on tomatillo in Keedysville. Around ten *C. virescens* and *C. subflexa* larvae sampled from each of tobacco, chickpeas, cotton and tomatillo plots in 2019 – 2021 were chosen for species identification in the preliminary experiments. The results verified that *C. subflexa* larvae were never collected in tobacco, chickpeas, and cotton, as expected, due to their specialization on *Physalis* hosts, nor were *C. virescens* larvae collected from tomatillo.

To further validate our findings of host plant use of *C. virescens* and *C. subflexa*, the number of larvae sampled per unit of time was also measured as an infestation indicator (Appendix Table 4.2 and 4.3). Overall, the average time spent sampling tobacco, tomatillo, chickpeas and cotton plots was 15.62, 40.42, 21.80, 35.20 minutes with standard deviations of 7.21, 14.63, 11.02, 14.40 minutes, respectively. As expected, more time was spent on sampling tomatillo and cotton plots due to the special sampling method; considering their canopy shape and the places *Chloridea* larvae live (usually in the fruits), it is laborious to search the whole plant. Because of this, the protocol for larval collection in cotton was slightly different in 2020: in the first 4 weeks, we used visual observations on every plant to document and collect each *Chloridea* larva, and then recorded the plant each larva came from. Later, as the plants grew up and no larvae were found, the sampling method was changed. Therefore, more time was spent on sampling cotton in the first half season of 2020. Besides plant canopy shape, many other factors could influence sampling time, e.g. the sampling participant and their familiarity with the protocol

(usually 2-3 students worked in the trap and field collection and they helped each other), the numbers and sizes of the larvae, the place where they hid, etc.

In 2020, the average numbers of larvae collected per unit time of *C. virescens* in tobacco, chickpeas and cotton were 0.14, 0.10 and 0 larvae/min with the order tobacco > chickpeas > cotton, and that of *C. subflexa* was 0.18 larvae/min. In 2021, the order of average larvae/min was changed to chickpeas (0.58 larvae/min) > tobacco (0.25 larvae/min) > cotton (0.01 larvae/min), and that of *C. subflexa* was 0.19 larvae/min. The results are consistent with previous measures of infestation rate, demonstrating that the number of larvae infesting each plant or plant size and shape did not strongly impact our ability to quantify infestation rate.

4.3.2 Trap collection

In total, 22 and 57 *C. virescens* and *C. subflexa* males were collected by pheromone-baited trap in 2020 and 2021, respectively (Figure 4.4). Only three *C. subflexa* were trapped in 2021, all from tomatillo traps. We did not collect any *C. subflexa* from the Upper Marlboro farm, and only one was trapped in the Beltsville farm in 2021. Most *C. virescens* (44 of 54) were trapped from the Upper Marlboro farm, and most (31 of 54) were from traps placed in tobacco fields in 2021. A Chi-squared test sought to test whether the trapped males of each species were equally distributed between traps placed in tobacco and tomatillo. For each year, our results showed modest differences in the distribution of trapped males of each *Chloridea* species among crops, but the differences were not significant at an α-level of 0.05 (χ^2_{2020}= 3.1145, p-value $_{2020}$ = 0.0776; χ^2_{2021}= 4.3542, p-value $_{2021}$ = 0.1134). Interestingly, we collected 11 and three *C. subflexa* from tomatillo traps with *C. virescens* lures in 2020

and 2021, respectively, while only one *C. subflexa* was collected from tobacco traps with *C. virescens* lures in 2020. Cross-attraction between *C. virescens* and *C. subflexa* (e.g. *C. subflexa* males captured in *C. virescens* female-baited traps) is not common in nature and few were observed in other field collections, but with live virgin females (Groot et al, 2009). In our study, a few *C. subflexa* males (11 of 22 in 2020; 3 of 57 in 2021) were attracted by *C. virescens* female pheromones in the field of their preferred host plant, tomatillo. Although the pheromone blends of *C. virescens* and *C. subflexa* are different, some of their pheromone components are similar but are present at different ratios. Taking this into consideration, a possible explanation of our results is that preferred host plant (tomatillo) odors could synergize with these pheromone components to attract *C. subflexa* males. An alternative hypothesis is that the special PRs of these trapped *C. subflexa* males enable them to respond to *C. virescens* female pheromones, following that the PRs should be different from those of wild *C. subflexa.*

4.3.3 Gene predictions and manual curations

In total, 15,098 protein coding genes were predicted in the genome of *C. subflexa* (Appendix Table 4.5). The average lengths of gene, exon, and intron regions were 8,308 bp, 249 bp and 1,226 bp, respectively. A BUSCO (version 4.1.2) analysis indicated that 1,234 (90.3%) of the 1,367 conserved insect genes were successfully annotated, including 1,227 single-copy genes and 7 duplicates (Table 4.1). Of the large numbers of gaps and ambiguous regions (180,381,610 bp Ns accounting for 38.69% of the assembly size) in the *C. subflexa* genome assembly, only 24,624,692 bp (5.28%) were identified as repeats. However, considering the difficulties in

assembling repeats using Illumina short reads, the gaps should contain most repeats. The three most abundant classes of repeat sequences included unclassified (4.33%), simple repeats (0.49%) and LINEs (0.20%), respectively (Appendix Table 4.5). The high proportion of unclassified elements may be due to the lack of studies on the repeat sequences of Noctuid species.

Pheromone receptors OR6, OR13, OR14 and OR16 were identified in the reference genomes of *C. virescens* and *C. subflexa* with the help of their mRNAs (Figure 4.5). Previous QTL studies showed that OR6 and OR14-16 are associated with species-specific male responses to female pheromone blends in *C. virescens* and *C. subflexa*, and they appear to be in close physical proximity within one chromosome (Gould et al, 2010). OR6 and OR14-16 were identified on a single scaffold in the *C. virescens* genome, but on different scaffolds in *C. subflexa* due to assembly fragmentation (Figure 4.5).

4.3.4 Examination of nucleotide diversity within ORs

Genetic diversity $\boldsymbol{\pi}$ (Nei, 1987) were examined for the coding regions (CDS) of OR6, OR13, OR14, OR16 and Orco per species (Table 4.2). Compared to other PRs, OR6 showed low levels of genetic diversity in both *C. virescens* and *C. subflexa*. Our results are consistent with previous evolutionary analysis in Cao et al (2021), which suggests that OR6 has the highest ω value in the PRs and is more likely to have undergone functional differentiation within the Heliothinae.

The ORs of *C. subflexa* have fewer variant sites than those of *C. virescens* (Table 4.2). A possible biological reason of the difference is that *C. virescens* is a generalist, while *C. subflexa* is a specialist, having more isolated populations across

North America (Groot et al, 2011). Less gene flow between populations could be a reason for fewer variant sites in *C. subflexa.* This low genetic variation could also be due to a potential technical limitation: the poor quality of the reference genome of *C. subflexa* (e.g. high repeat content). Although gaps only exist in the introns of ORs in *C. subflexa*, they would still affect read alignment in the coding regions near the gaps (Appendix Figure 4.2). As a result, there were no reads mapped to some regions of the genes, even the coding regions, in some individuals. Particularly, one region of the OR14 CDS had no read alignment for all the individuals in *C. subflexa*, meaning that the region could not be used in these analyses.

4.3.5 Comparison of amino acid variants and their distributions across *Chloridea* ORs

Overall, in *C. virescens* the protein sequences of OR6, OR13, OR14 and OR16 contain one, three, 21 and six variant sites, respectively, while in *C. subflexa* the protein sequences of OR6, OR14 and OR16 contain one, nine and four, respectively. No amino acid variations were detected in *C. subflexa* OR13. The allele frequencies of nonsynonymous SNPs of each OR were compared between populations (*C. virescens* trapped from tobacco vs. tomatillo, *C. subflexa* trapped from tomatillo vs. wild *C. subflexa* collected from tomatillo plots), and Fisher's exact tests with Bonferroni corrections for multiple comparisons showed that there were no significant differences between populations (Appendix Table 4.6 and 4.7).

We then examined how the amino acid variations were distributed within the receptors (Figure 4.6 and 4.7). Transmembrane topologies predicted by TMHMM (Krogh et al, 2001) for each OR divided amino acid residues into three receptor

regions: the intracellular loops (IL), the transmembrane domains (TMD; seven per OR) and the extracellular loops (EL). Most amino acid variants (29 of 45) were distributed in the intracellular loops (IL). Only *C. subflexa* OR6 showed a high level of variation in the EL (Appendix Figure 4.3). Statistically, no differences between the expected frequencies and the observed frequencies of amino acid variation distribution could be detected across the three regions using chi-squared tests at an α-level of 0.05 (Appendix Table 4.8).

Notably, OR6 only has one amino acid variant in *C. virescens* and *C. subflexa*. The amino acid variant was located in either the IL (*C. virescens*) or the EL (*C. subflexa*). In *Ostrinia furnacalis* and *Ostrinia nubilalis*, a single amino acid mutation in TMD3 could change the ligand specificity of OR3 (Leary et al., 2012). In two Noctuidae species, *H. armigera* and *H. assulta,* two point-mutations in TMD4 and TMD6 could shift the ligand selectivity of OR14b (Yang et al., 2017). In addition, recent studies identified a key site in TMD5 causing functional changes between CvirOR6 and CsubOR6 (Cao et al, 2021). Therefore, given these studies and our observation of a lack of variation in the TMD region, we speculate that TMD plays an important role in functional differentiation of PRs.

4.3.6 Analysis of Molecular Variance (AMOVA)

In the 65 complete ORs (of 80) found in the *C. virescens* genome, AMOVA tests were performed to determine whether host plant choice was associated with genetic variation in *C. virescens.* The results showed that most variation was either distributed within samples or between samples within host-plant groups rather than between host-plant groups (Table 4.4). The random permutation tests between plant

groups were not significant for all the ORs with a Bonferroni corrected $\alpha = 0.000769$ (0.05/65 compared pairs). However, considering Bonferroni correction is a very conservative method that controls for the probability of type I error, we found that the random permutation tests were significant at an α-level of 0.05 in OR6, OR55, OR66 and OR78 genes (p-values: OR6 = 0.03497, OR55 = 0.03796, OR66 = 0.00799, OR78 = 0.04296), indicating that host plant associated population differentiation existed in these genes.

F_{st} values (Weir and Cockerham, 1984) were used to estimate genetic divergence of OR6 between the *C. virescens* populations, and we found that the highly divergent sites (top 5% SNPs) were all in introns (Figure 4.8). As introns play important roles in gene expression and regulation, these highly divergent sites may play a potential role in the expression of OR6. For example, a novel splice site (AG) was introduced by two variants in the introns (positions 27,172 and 28,627 of scaffold NWSH01000007.1), which have a potential to cause alternative splicing in the transcription. A Chi-squared test demonstrated that the distribution of these highly divergent sites was not biased within OR6 (Chi-square = 2.59, p-value = 0.11). Linkage disequilibrium (LD) values showed that these highly divergent sites were not linked (Appendix Figure 4.4). They are only inherited together when they are very close, but not across the whole gene.

The trapped *C. subflexa* were attracted by baited *C. virescens* lures in tomatillo plots, and we expect their PRs to be different from the PRs of wild *C. subflexa*, which were collected from the tomatillo plots and were reared to adults in the lab. A comparison was made between the two *C. subflexa* populations to test the

hypothesis, and the results showed that only a small part of the variation (OR13 = 4.7%, OR6, OR14 and OR16 near to 0) could be explained by the collection methods (*C. virescens* lures), while most of the variation was explained by differences between and within samples, regardless of the collection method (Table 4.5).

4.3.7 Selection on PRs in wild *C. virescens* and *C. subflexa*

Tajima's D (Tajima, 1989) was calculated to test whether the observed mutations of each OR were caused by non-random processes (e.g. natural selection or population size change). All ORs showed positive Tajima's D values (Table 4.3), which indicate low levels of low and high frequency alleles across these genes. An excess of intermediate frequency alleles could be produced by both genetic drift and selection. Therefore, we examined the probability that such Tajima's D values could be reached under conditions of neutrality (no selection) using coalescent simulations (Table 4.3). Simulations demonstrated that the probability was low (p-value = 0.080 - 0.104) with population size 5,000 - 20,000 (estimated from Blanco et al, 2007). Except for balancing selection, sudden population contraction could be another explanation for positive Tajima's D, so we simulated a population whose size decreased from N_0 = 20,000 to 30% at half-time (t = 0.5), and the results demonstrated that the probability of observing our result given this process was also very low (p-value = 0.028).

To further evaluate the evolutionary forces acting on the PRs, we estimated the rates of nonsynonymous to synonymous substitutions (dN/dS or ω) for OR6, OR13, OR14, OR16 and Orco orthologs with haplotypes predicted from the *C.*

virescens and *C. subflexa* field populations (Figure 4.9). The ω values of all the five OR orthologs were much smaller than 1, indicating that they are under strong purifying selection. The ω for Orco orthologs was the lowest ($\omega = 0.003$) as expected, given the high functional conservation of Orco genes (Vosshall and Hansson, 2011). The ω for OR6 orthologs was the highest ($\omega = 0.193$), which is consistent with our expectation and Cao et al (2021), considering the functional diversification of OR6 in *C. virescens* and *C. subflexa*. However, in Cao et al (2021), the PRs of *C. virescens* and *C. subflexa* were amplified from a lab colony that was reared with artificial conditions, while in our study the haplotypes were predicted from wild populations. The lab colonies that were maintained at North Carolina State University (Wang et al, 2011; Cao et al, 2021) are quite different from the wild populations, and for all ORs except HsOR6, we did not observe their haplotypes in our field collected samples. The identities of mRNA alignment of the haplotypes having the highest probabilities in the field populations against the haplotypes in Cao et al (2021) are: HvOrco: 97.39%, HvOR6: 97.91%, HvOR13: 99.37%, HvOR14: 97.51%, HvOR16: 96.30%, HsOrco: 99.65%, HsOR6: 99.92%, HsOR13: 99.92%, HsOR14: 99.92%, HsOR16: 99.37%.

Previous studies showed that the majority of alleles present in field-collected *C. virescens* populations were low frequency alleles (Fritz et al, 2016), and we can see purifying selection acting on the ORs to remove low frequency alleles. Our results support a combination of the signal-response coevolution (Phelan, 1992; Lofstedt, 1993) and Phelan's asymmetric tracking (Phelan, 1997) theories existing in nature and for these species. In sex pheromone communications, males are required to have

broader responses to both wild-type and mutant female pheromones, so maintaining genetic diversity at the pheromone receptor genes is necessary for persistence of the species. However, males that fail to respond to female signals will not be able to pass its genetic information onto its offspring, resulting in purifying selection acting on the PRs to remove low frequency alleles.

4.4 Conclusion

In summary, in this chapter we characterized host plant use by *C. virescens* and *C. subflexa* in central Maryland over 8-week field collections in 2020 and 2021. Sequencing and examination of nucleotide diversity and tests of selection on a subset of those field-collected samples indicated that OR6 was under very strong purifying selection in both *C. virescens* and *C. subflexa*. AMOVA tests suggested that in *C. virescens*, host plant associated genetic differentiation existed in genes OR6, OR55, OR66 and OR78, and further genetic divergence analysis showed that the highly divergent sites were all in introns of OR6. Because few previous studies have been done on genes OR55, OR66 and OR78, further analysis would focus on these genes in *C. virescens*.

Tables: Chapter 4

Table 4.1. Structural annotation statistics of *C. subflexa*.

	Statistics
Protein coding genes	15,098
Genes % of genome	27%
Mean gene length (bp)	8,308
Exons	95,350
Exons % of genome	5%
Mean exon length (bp)	249
Introns	80,252
Introns % of genome	22%
Mean introns length (bp)	1,266
BUSCO insecta_odb10	
complete (C)	1234 (of 1367, 90.3%)
single copy (S)	1227 (89.8%)
duplicted (D)	7 (0.5%)
fragmented (F)	55 (4.0%)
missing (M)	78 (5.7%)

Table 4.2. Genetic diversity (π), nucleotide variant sites and Tajima's D values of OR6, OR13, OR14, OR16 and Orco in *C. virescens* and *C. subflexa*.

	***C. virescens* (n = 18)**			***C. subflexa* (n = 20)**		
ORs	**Pi_CDS**	**Variant sites_CDS**	**Tajima's D_whole gene**	**Pi_CDS**	**Variant sites_CDS**	**Tajima's D_whole gene**
Orco	0.386	54	1.81	0.300	18	2.00
OR6	0.323	41	1.45	0.348	10	1.47
OR13	0.358	18	1.58	0.224	2	0.33
OR14	0.396	51	1.54	0.424	16	2.70
OR16	0.365	51	1.86	0.405	18	2.39

Table 4.3. Probabilities of observing Tajima's D greater than or equal to the empirical threshold 1.45 in coalescent-based simulations, and parameters used in 1,000 simulations of each parameter combination.

N_0	**Total Sims**	**Population contraction**	**Mean Tajima's D**	**Prob Tajima's D >= 1.45**
20,000	1,000	30% at time = 0.5	-0.327	0.028
20,000	1,000	No	-0.033	0.083
15,000	1,000	No	-0.034	0.080
10,000	1,000	No	-0.043	0.095
5,000	1,000	No	0.00088	0.104

Table 4.4. Analysis of molecular variance (AMOVA) for *C. virescens* populations trapped from tobacco vs. tomatillo plots, with applying a Bonferroni corrected alpha level (corrected $\alpha = 0.05/65$ compared pairs = 0.000769).

	Variations %				p-value
	Between plant %	**Between samples Within plant %**	**Within samples %**	**Total %**	**Between plants**
OR1	2.64	58.1	39.26	100	0.15584
OR4	0.42	18.02	81.56	100	0.40959
OR6*	3.48	39.86	56.66	100	0.03497*
OR7	-0.3	30.48	69.82	100	0.50150
OR9	-0.66	27.43	73.24	100	0.57542
OR10	-0.21	40.01	60.2	100	0.47153
OR12	-2.54	51.7	50.84	100	0.78921
OR13	-2.87	23.47	79.4	100	0.90010
OR14	1.19	-11.64	110.45	100	0.14585
OR16	2.47	42.21	55.32	100	0.12987
OR18	-1.75	20.78	80.96	100	0.80120
OR19	1.07	25.23	73.7	100	0.32867
OR20	1.66	50.13	48.21	100	0.21878
OR21	5.22	37.91	56.87	100	0.11089
OR23	0.53	38.08	61.38	100	0.33167
OR24	-2.45	2.87	99.57	100	0.64236
OR26	-4.44	62.42	42.02	100	0.91808
OR27	4.59	73.52	21.89	100	0.18681
OR28	-0.67	46.19	54.48	100	0.56344
OR29	0.33	31.13	68.54	100	0.39061
OR30	-0.9	15.53	85.37	100	0.45155
OR35	-2.09	41.4	60.69	100	0.76124
OR36	-0.42	41.8	58.63	100	0.52448
OR37	-3.03	15.96	87.06	100	0.93107
OR38	-1.87	14.31	87.56	100	0.75524
OR41	0.96	47.1	51.94	100	0.28272

*p-value < 0.05; **p-value < 0.01

Table 4.4. Continued.

	Variations %				p-value
	Between plant %	**Between samples Within plant %**	**Within samples %**	**Total %**	**Between plants**
OR41-like	1.2	48.47	50.33	100	0.31469
OR42	-1.23	-6.97	108.21	100	0.75125
OR44	0.98	26.19	72.83	100	0.23177
OR52	3.12	73.87	23.02	100	0.18581
OR53	0.6	40.25	59.14	100	0.28372
OR54	-0.2	20.93	79.27	100	0.45754
OR55*	8.25	7.86	83.89	100	0.03796*
OR56	-0.64	37.5	63.13	100	0.49051
OR57	-1.35	34.31	67.04	100	0.70629
OR58	-2.21	23.8	78.41	100	0.85814
OR59	-0.71	8.27	92.43	100	0.53546
OR60	-5.36	60.87	44.49	100	0.98601
OR62	-3.29	15.82	87.47	100	0.77123
OR63	-0.09	31.17	68.92	100	0.45954
OR64	-1.88	28.75	73.13	100	0.73626
OR66**	6.31	25.72	67.97	100	0.00799**
OR67	3.94	16.72	79.35	100	0.06893
OR68	-0.32	44.16	56.16	100	0.50050
OR69	0.48	20.05	79.47	100	0.32867
OR71	-0.13	66.53	33.6	100	0.49451
OR72	1.33	29.42	69.25	100	0.21279
OR78*	6.97	39.44	53.59	100	0.04296*
OR80	-0.25	25.9	74.35	100	0.47053
OR81	-4.12	37.58	66.54	100	0.99201
OR82	-1.19	58.66	42.53	100	0.59341
OR84	1.7	56.3	42	100	0.24675
OR85	-1.03	68.18	32.84	100	0.54845
OR86	0.66	46.37	52.97	100	0.35864
OR87	3.96	67.57	28.46	100	0.15385

*p-value < 0.05; **p-value < 0.01

Table 4.4. Continued.

	Variations %				p-value
	Between plant %	**Between samples Within plant %**	**Within samples %**	**Total %**	**Between plants**
OR88	-4.9	67.13	37.78	100	0.92108
OR89	-1.1	30.89	70.21	100	0.64735
OR90	-3.83	64.49	39.34	100	0.88012
OR91	-0.19	45.95	54.24	100	0.45554
OR92	-2.16	77.34	24.82	100	0.53746
OR93	-0.3	29.49	70.81	100	0.37862
OR94	-3.14	44.39	58.74	100	0.62138
OR95	1.09	-24.34	123.25	100	0.24076
OR96	-3.04	77.66	25.37	100	0.71928
OR97	1.16	59.07	39.77	100	0.26673

*p-value < 0.05; **p-value < 0.01

Table 4.5. Analysis of molecular variance (AMOVA) for *C. subflexa* populations*, with applying a Bonferroni corrected alpha level (corrected $\alpha = 0.05/4$ compared pairs = 0.0125).

	Variations %				p-value
	Between CollectionWay %	**Between samples Within CollectionWay %**	**Within samples %**	**Total %**	**Between CollectionWays**
OR6	-3.14	15.25	87.89	100	0.9091
OR13	4.7	33.25	62.04	100	0.2078
OR14	-3.39	-15.14	118.53	100	0.9131
OR16	-1.29	4.76	96.53	100	0.5944

***** Two populations: 10 *C. subflexa* collected from pheromone traps in tomatillo plots vs. 10 wild *C. subflexa* that were collected from tomatillo plots as larvae and reared to adults in the lab.

Figures: Chapter 4

Figure 4.1. Pheromone-baited trap for *Chloridea* species set up in a tomatillo field at Upper Marlboro farm.

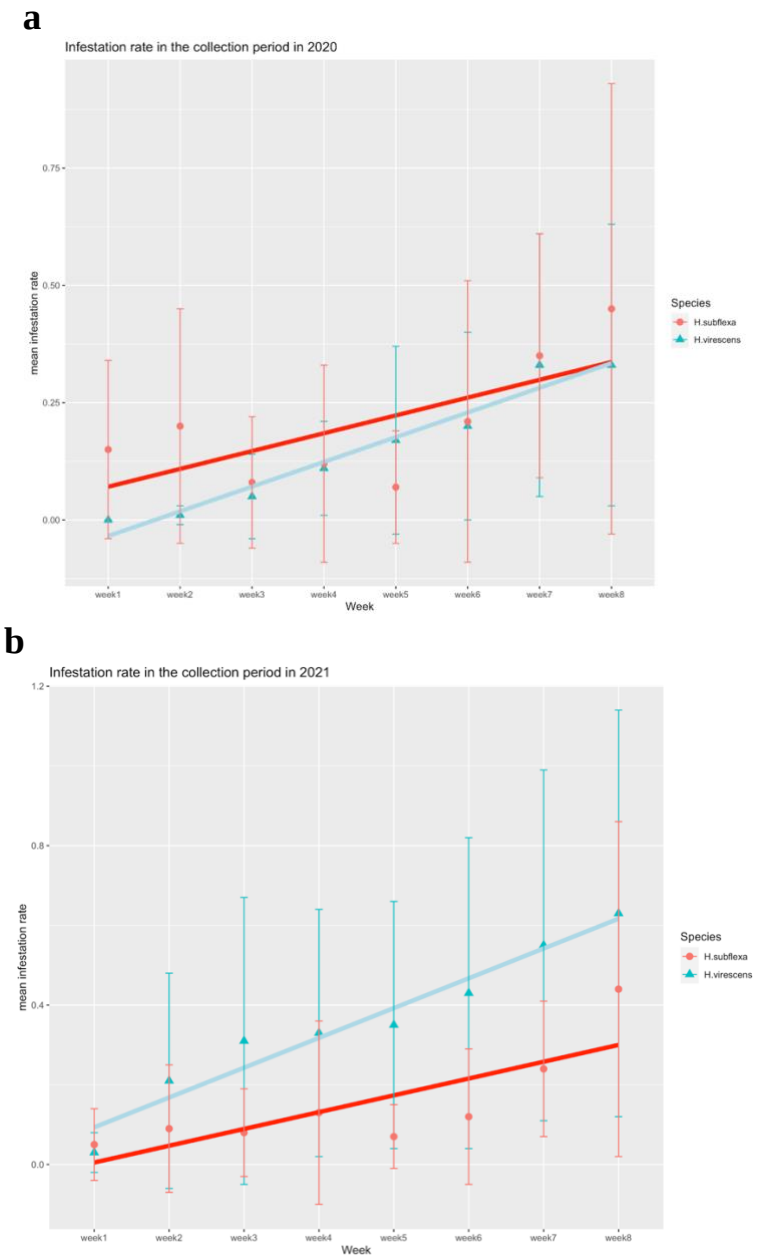


Figure 4.2. Average infestation rates of *C. virescens* (blue) and *C. subflexa* (red) in the 8-week collection period in 2020 (a) and 2021 (b). Infestation rate: number of infested plants / number of sampled plants * 100.

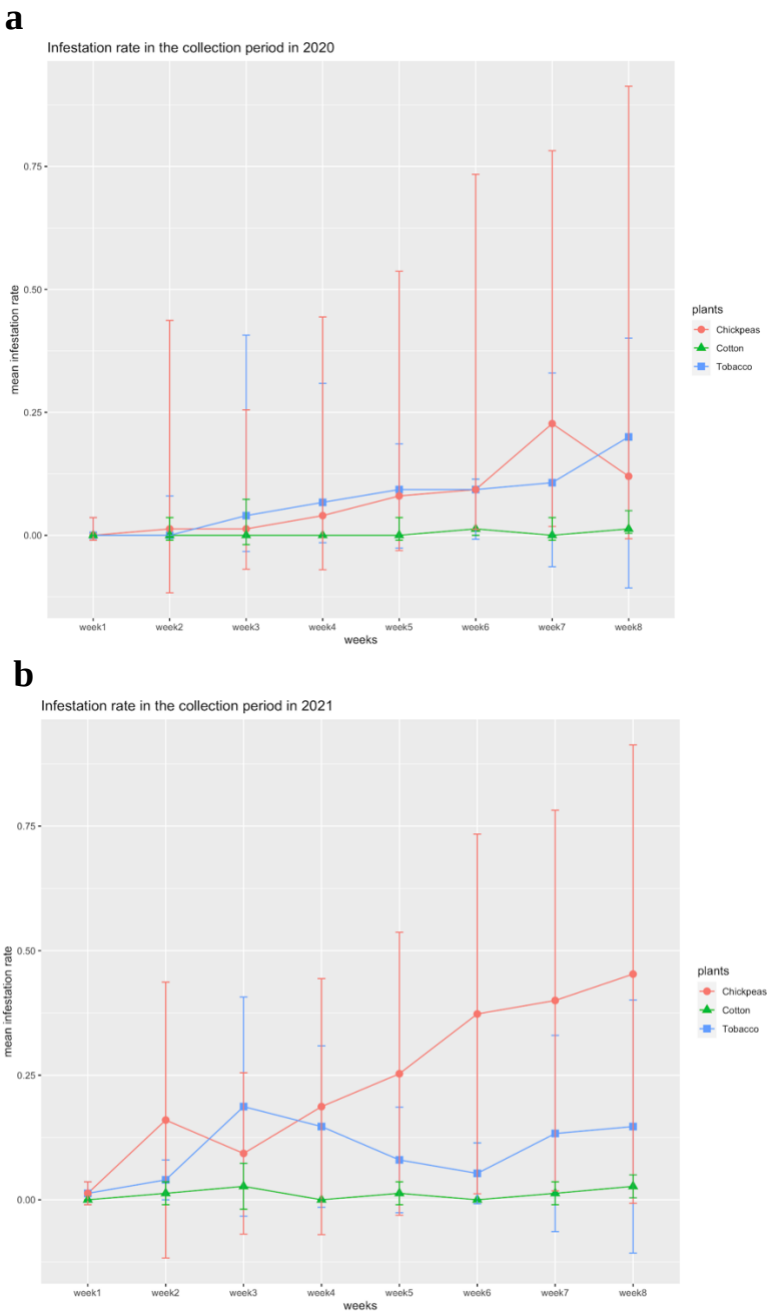


Figure 4.3. Average infestation rates of *C. virescens* in host plants tobacco (blue), chickpeas (red) and cotton (green) in the 8-week collection period in 2020 (a) and 2021 (b). Infestation rate: number of infested / number of sampled plants * 100.

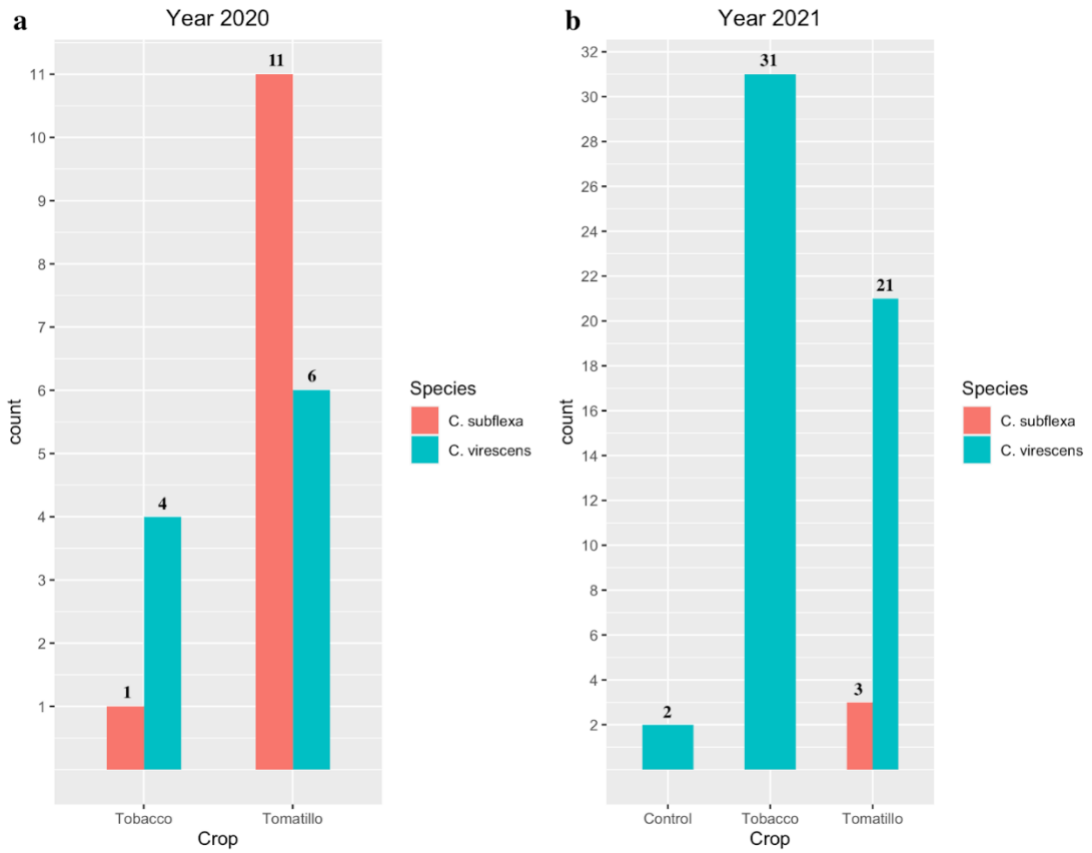
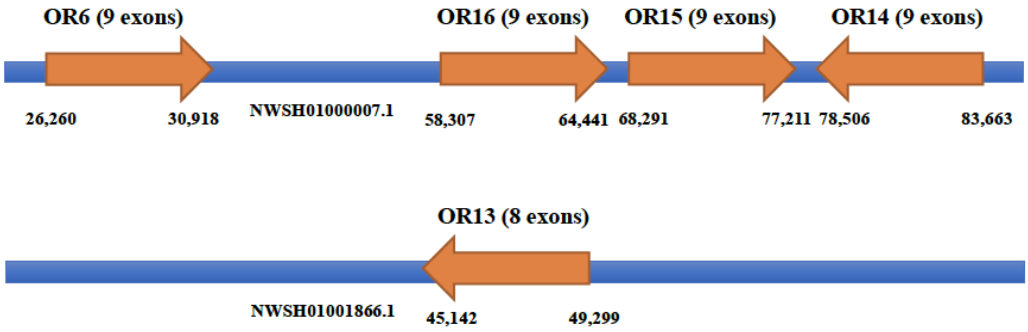


Figure 4.4. Number of trapped *C.virescens* (blue) and *C. subflexa* (red) males by crop in (a) 2020 and (b) 2021.

A



B

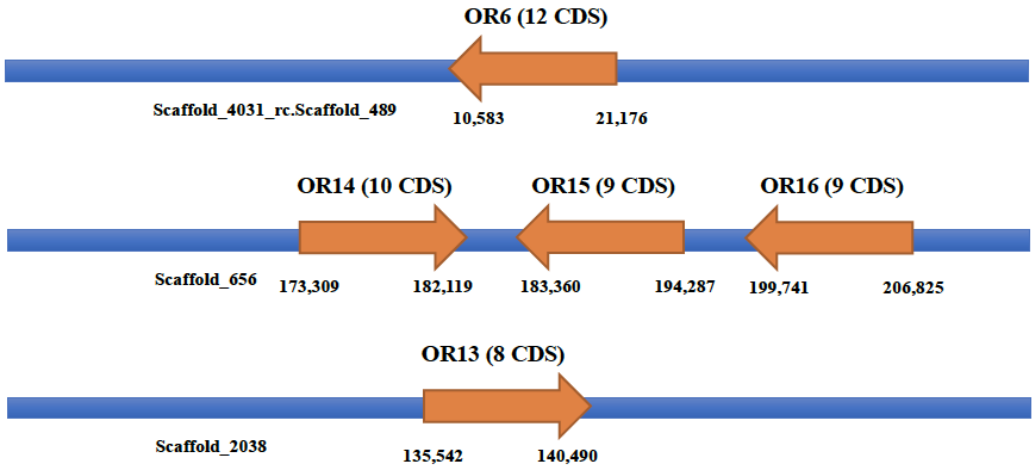


Figure 4.5. Locations of OR6, OR13, OR14-16 in the reference genomes of *C. virescens* (A; GenBank assembly accession: GCA_002382865.1) and *C. subflexa* (B; GenBank assembly accession: GCA_022398575.1).

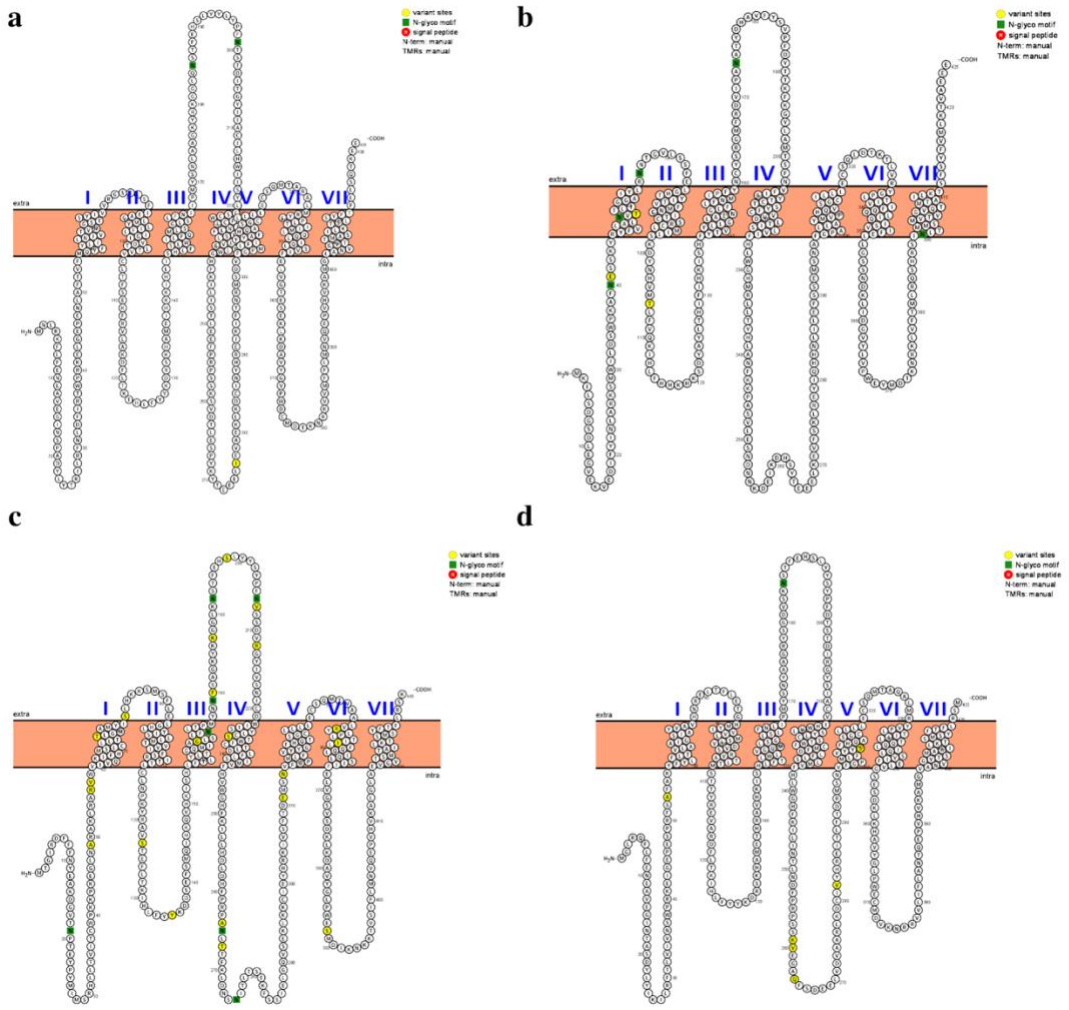


Figure 4.6. Predicted transmembrane topologies of *C. virescens* OR6 (a), OR13 (b), OR14 (c) and OR16 (d), with variant sites within the species highlighted in yellow.

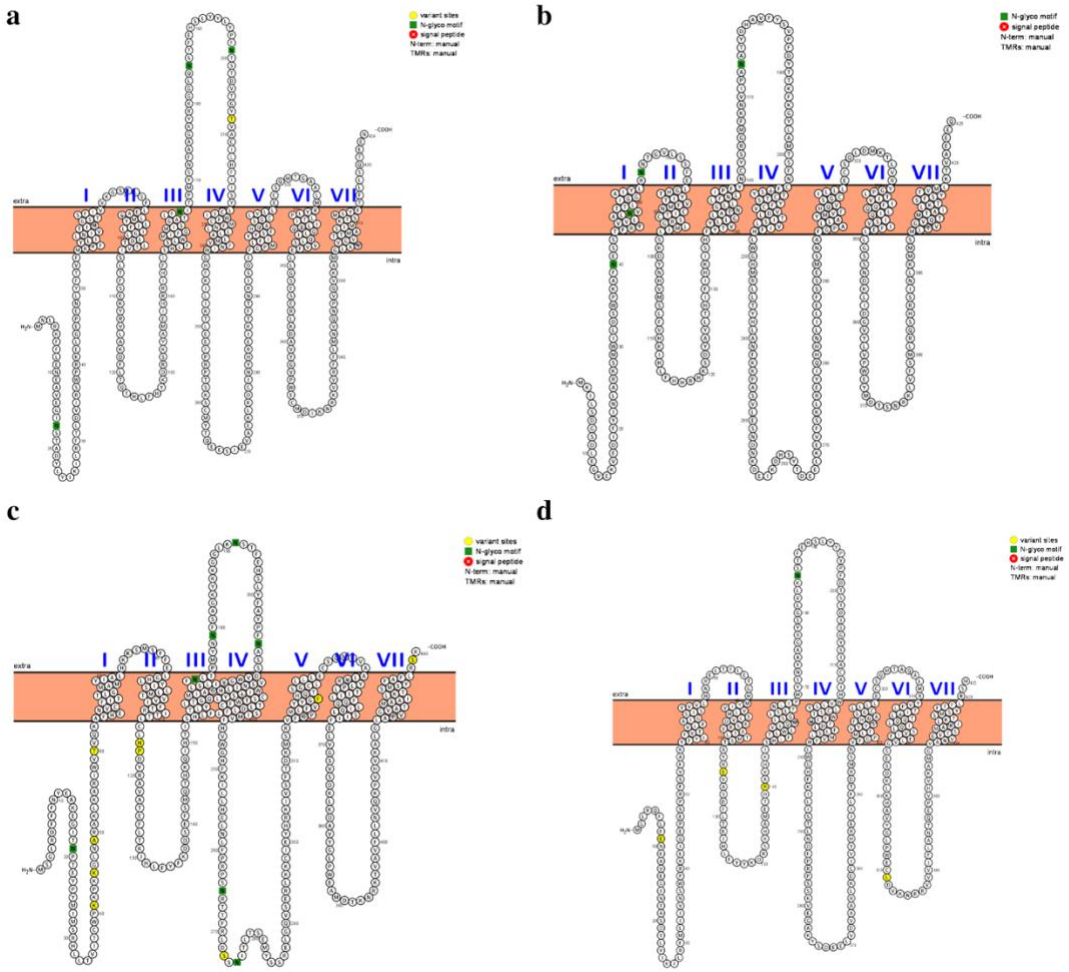


Figure 4.7. Predicted transmembrane topologies of *C. subflexa* OR6 (a), OR13 (b), OR14 (c) and OR16 (d), with variant sites within the species highlighted in yellow. No variant sites were detected in OR13.

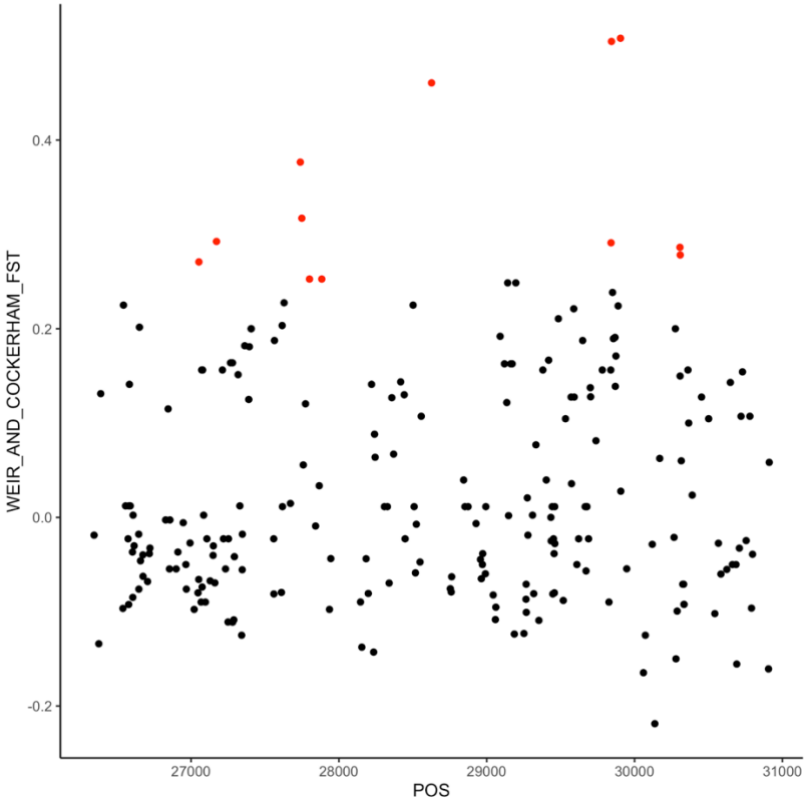


Figure 4.8. Weir and Cockerham Fst values of OR6 between *C. virescens* populations trapped from tobacco and tomatillo traps. Each dot represents a SNP and the x axis represents its position. SNPs with the highest 5% Fst values shown in red.

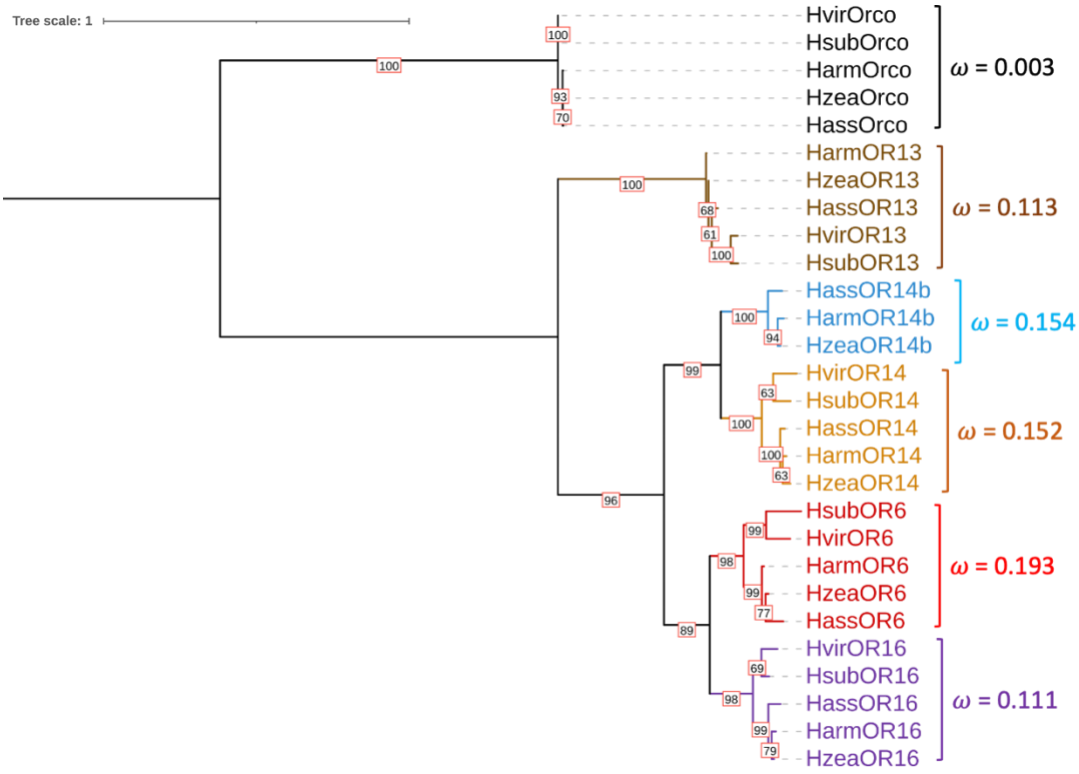


Figure 4.9. Phylogenetic tree of OR6, OR13, OR14, OR16 and Orco in Heliothine moths. Tree scale represents the number of amino acid substitutions per site. Bootstrap values were based on 1,000 replicates. Hvir: *C. virescens*, Hsub: *C. subflexa*, Harm: *H. armigera*, Hzea: *H. zea*, Hass: *H. assulta.*

Chapter 5: Final conclusions and future directions

The insect chemosensory system is exquisitely tuned for detection of chemical signals in the environment. These chemical signals are involved in the process of host plant detection, species recognition, mate recognition and discrimination, predator avoidance, social communication, etc., which are important for insect fitness (Wilson, 1965; Kawano et al, 1999; Johansson and Jones, 2007). Natural selection acts upon the chemosensory system to refine a species ability to detect volatiles in ways that improve insect fitness and reinforce species boundaries (Smadja and Butlin, 2009; Vieira et al, 2007; Vieira and Rozas, 2011). Studying the evolution of insect chemosensory systems requires available genomic resources. My dissertation focused on the development of tools to investigate the evolution of chemosensory gene families involved in host plant and mate recognition in Heliothine moths, with special focus on two closely-related lepidopteran species, *C. virescens* and *C. subflexa*. I then applied these tools to examine inter- and intraspecific diversity of two chemosensory gene families: the ORs and OBPs.

In Chapter 2, I presented two Heliothine genome assemblies: a novel *C. subflexa* short read assembly and an improved, highly contiguous *C. virescens* assembly. These were used along with the existing *Helicoverpa zea* and *H. armigera* assemblies to demonstrate that reference-assisted assembly, has the potential to enhance the contiguity and completeness of fragmented assemblies, at the expense of additional gap opening. This was particularly true for *C. subflexa*, which I documented to have the largest genome-size of known Heliothine genome sizes, to date. Kmer analysis revealed that this genome size expansion partly resulted from an

increase in the percentage of the genome comprised of repetitive elements, which was unique to *C. subflexa*. The gaps opened in the *C. subflexa* genome are likely to contain repetitive elements but could also contain chemosensory genes. Chemosensory gene family evolution occurs through a process of gene duplication and diversification (Sánchez-Gracia et al, 2009), which I speculate would impact ascertainment of the full chemosensory suite in my assembly. While my novel *C. subflexa* genome will likely serve as a critical resource for investigation of some gene families, future work should be directed at genome improvement to ascertain the full suite of *C. subflexa* ORs and OBPs.

Chapter 3 focused on manual curation of two chemosensory gene families in the *C. virescens* genome, but not the *C. subflexa* genome due to its fragmentation. These two gene families, the ORs and OBPs, play evolutionarily important roles in detection of odors for mate recognition and host plant identification. I manually annotated Orco, 80 ORs, and 49 OBPs, which were classified as either Classic (n = 34), Minus-C (n = 8), and Plus-C OBPs (n = 7). Phylogenetic analyses identified potential gene duplication and loss of OR and OBP gene families among members of the Heliothinae, and two novel gene diversification events observed in the "Minus-C" group of *B. mori* and the Heliothine moths. Gene diversification events were also found in OR6 and OR14 of the PR clade. I postulate that diversification events occurring in these gene families may be related to differences in volatile sensation and olfactory behaviors among Heliothine species. Future studies could focus on verifying the results, for example, by knocking out the correspond genes in the lab lines with CRISPR and studying the behavior of mutants.

In Chapter 4, I documented the phenology and host plant use by *C. virescens* and *C. subflexa* at three University research farms in the state of MD. Samples collected from these field sites were sequenced and examined for inter- and intraspecific evidence of selection at chemosensory genes, with special focus on pheromone receptors. OR6, an important pheromone receptor, was under very strong purifying selection in both *C. virescens* and *C. subflexa* according to dN/dS. Moreover, AMOVA tests suggested that in *C. virescens*, host plant associated population differentiation existed in genes OR6, OR55, OR66 and OR78. Future work could focus on examining the role of single nucleotide polymorphism and structural variants in these genes in the wild *C. virescens* populations and the functional role of these genes in odorant detection. Overall, the studies conducted here contribute to our understanding of insect chemosensory systems and provide important insight into the genetic mechanisms underlying host plant specialization and sexual communication among lepidopteran insects.

Appendix

Additional figures and tables for each chapter are given below.

Chapter 2 Tables

Table 2.1. The summary of insect genome assemblies across insect orders (Last updated: 2020-07-16)

Order	**Assemblies**
Archaeognatha	1
Blattodea	6
Coleoptera	34
Dermaptera	1
Diptera	164
Ephemeroptera	3
Hemiptera	46
Hymenoptera	131
Lepidoptera	87
Odonata	3
Orthoptera	3
Phasmatodae	4
Phthiraptera	1
Plecoptera	3
Siphonaptera	1
Strepsiptera	1
Thysanoptera	2
Trichoptera	6
Total	497

Table 2.2. The number of Lepidopteran genome assemblies (Triant et al, 2018; Last updated: 2020-07-16)

Superfamily	Families	Species	Assemblies	Proportional assembly effort
Bombycoidea	10	4723	6	0.127%
Lasiocampoidea	1	1952	1	0.051%
Geometroidea	5	23748	1	0.004%
Noctuoidea	6	42407	13	0.031%
Drepanoidea	3	672	0	0
Mimallonoidea	1	194	0	0
Pyraloidea	2	15576	6	0.039%
Papilionoidea	7	18768	52	0.277%
Gelechioidea	16	18489	2	0.011%
Thyridoidea	1	940	0	0
Pterophoroidea	1	1318	0	0
Alucitoidea	2	235	0	0
Copromorphoidea	2	326	0	0
Epermenioidea	1	126	0	0
Hyblaeoidea	1	19	0	0
Calliduloidea	1	49	0	0
Cossoidea	6	2881	0	0
Zygaenoidea	12	3296	0	0
Tortricoidea	11	10387	2	0.019%
Urodoidea	1	65	0	0
Choreutoidea	1	406	0	0
Schreckensteinioidea	1	8	0	0

Superfamily	Families	Species	Assemblies	Proportional assembly effort
Immoidea	1	245	0	0
Millieriidae	1	4	0	0
Galacticoidea	1	19	0	0
Douglasiidae	3	29	0	0
Yponomeutoidea	10	1756	1	0.057%
Gracillarioidea	3	2216	1	0.045%
Tineoidea	5	3823	1	0.026%
Palaephatoidea	1	57	0	0
Tischerioidea	1	110	0	0
Andesianoidea	1	3	0	0
Adeloidea	6	582	1	0.172%
Nepticuloidea	2	1011	0	0
Hepialoidea	2	636	0	0
Lophocoronoidea	1	6	0	0
Eriocranioidea	1	29	0	0
Neopseustoidea	3	9	0	0
Heterobathmioidea	1	3	0	0
Agathiphagoidea	1	2	0	0
Micropterigoidea	1	160	0	0
Total	137	157285	87	.055%

Table 2.3. Genome size estimations of *C. virescens* and *C. subflexa*

Samples	PI-control	PI-sample	Estimated genome sizes
Hsub Female 1	1177	3226	479.65
Hsub Female 2	1156	3017	456.73
Hsub Female 3	1111	2926	460.89
Hsub Male 1	1141	3061	469.48
Hsub Male 2	1122	2985	465.57
Hsub Male 3	1167	3127	468.92
Hvir Female 1	1131	2482	384.04
Hvir Female 2	1113	2410	378.93
Hvir Female 3	1149	2522	384.12
Hvir Male 1	1052	2304	383.27
Hvir Male 2	1006	2193	381.49
Hvir Male 3	1127	2448	380.12

Table 2.4. Counts of Illumina reads used to produce the *C. subflexa* genome assembly before and after trimming.

	Before Trimming		**After Trimming**	
Libraries	Total sequences	%GC	Total sequences	%GC
MP_2kb_R1	54,598,005	39%	48,497,353	39%
MP_2kb_R2	54,598,005	39%	48,497,353	39%
MP_5kb_R1	42,472,942	40%	37,800,771	40%
MP_5kb_R2	42,472,942	39%	37,800,771	39%
PE1_R1	90,482,204	37%	90,434,531	36%
PE1_R2	90,482,204	36%	90,434,531	36%
PE2_R1	92,066,056	37%	92,020,691	36%
PE2_R2	92,066,056	36%	92,020,691	36%

Table 2.5. Blast results of partial cDNA sequences of *C. subflexa* OR6 (HM751833.1), OR14 (HM751835.1), OR15 (HM751836.1) and OR16 (HM751837.1) to the assembly

OR6_position	1-80	81-183	184-230	243-279	278-348	363-445
Scaffold_489_position	5398-5319	4935-4833	4753-4707	4559-4523	4108-4038	3899-3818
Identities (%)	100%	100%	100%	100%	100%	96.39%

OR14_position	1-44	45-321	322-432	430-666	665-747
Scaffold_656_position	1703309-173352	173687-173963	174468-174578	174732-174968	175196-175278
Identities (%)	100%	100%	100%	99.16%	100%

OR15_position	1-44	45-321	322-432	430-666
Scaffold_656_position	194222-194179	192996-192720	192346-192236	191286-191050
Identities (%)	100%	100%	100%	100%

OR16_position	1-86	87-189	189-285	284-441	441-489
Scaffold_656_position	203365-203280	202483-202381	202306-202210	201772-201615	199974-199926
Identities (%)	100%	100%	100%	99.37%	100%

Table 2.6. Mapping back rates for subsets of *C. virescens* raw reads from SRA accession: SRR5463746 and ddRAD-seq data from wild *C. virescens* against the published *C. virescens* assembly (GenBank assembly accession: GCA_002382865.1), a novel Dovetail *C. virescens* assembly and two Ragout extended assemblies.

Mapping rates	***C. virescens***	**Dovetail**	***C. vir_solid***	***C. vir_unsolid***
subset_1	87.43%	87.43%	87.43%	87.42%
subset_2	87.42%	87.42%	87.43%	87.42%
subset_3	87.45%	87.45%	87.46%	87.45%
subset_4	87.43%	87.43%	87.43%	87.42%
subset_5	87.41%	87.41%	87.41%	87.41%
Average	87.43%	87.43%	87.43%	87.42%
SD	0.01%	0.01%	0.02%	0.02%
LA2012_3	55.38%	55.40%	55.40%	55.13%
LA2012_4	54.73%	54.73%	54.75%	54.64%
LA2012_7	55.43%	55.45%	55.43%	55.15%
LA2012_8	55.29%	55.31%	55.31%	55.15%
LA2012_9	55.36%	55.39%	55.38%	55.22%
Average	55.24%	55.26%	55.25%	55.06%
SD	0.29%	0.30%	0.29%	0.24%

Table 2.7. Quality metrics for the published and improved *H. zea* genome assemblies

Statistics	***H. zea***	***H. zea_solid***	***H. zea_unsolid***
Assembly size (Mb)	341.15	371.29	375.54
Number of scaffolds	2,975	1,305	1,446
Max scaffold length (bp)	1,847,547	21,379,678	18,478,309
Min scaffold length (bp)	8,503	8,503	110
Mean scaffold length (bp)	114,671	284,510	259,712
N50 (bp)	201,477	6,100,690	8,865,798
N90 (bp)	52,272	95,860	90,526
N_count (bp)	34,741,529	64,879,699	69,137,956
BUSCO insecta_odb10			
complete (C)	1318 (of 1367, 96.4%)	1320 (of 1367, 96.6%)	1317 (of 1367, 96.3%)
single copy (S)	1305 (95.5%)	1305 (95.5%)	1300 (95.1%)
duplicated (D)	13 (1.0%)	15 (1.1%)	17 (1.2%)
fragmented (F)	13 (1.0%)	14 (1.0%)	14 (1.0%)
missing (M)	36 (2.6%)	33 (2.4%)	36 (2.7%)
BUSCO eukaryota_odb10			
complete (C)	240 (of 255, 94.2%)	244 (of 255, 95.7%)	243 (of 255, 95.3%)
single copy (S)	235 (92.2%)	239 (93.7%)	239 (93.7%)
duplicated (D)	5 (2.0%)	5 (2.0%)	4 (1.6%)
fragmented (F)	9 (3.5%)	6 (2.4%)	7 (2.7%)
missing (M)	6 (2.3%)	5 (1.9%)	5 (2.0%)

Table 2.8. Quality metrics for the published and improved *H. armigera* assemblies.

Statistics	***H. armigera***	***H. armigera_solid***	***H. armigera_unsolid***
Assembly size (Mb)	337.09	352.34	362.74
Number of scaffolds	998	594	795
Max scaffold length (bp)	6,146,627	24,694,291	15,550,744
Min scaffold length (bp)	8,841	10,004	107
Mean scaffold length (bp)	337,763	593,171	456,274
N50 (bp)	1,000,414	9,477,009	9,240,022
N90 (bp)	175,335	313,099	318,902
N_count (bp)	37,110,157	52,366,090	62,760,126
BUSCO insecta_odb10			
complete (C)	1332 (of 1367, 97.4%)	1332 (of 1367, 97.4%)	1328 (of 1367, 97.1%)
single copy (S)	1326 (97.0%)	1326 (97.0%)	1322 (96.7%)
duplicated (D)	6 (0.4%)	6 (0.4%)	6 (0.4%)
fragmented (F)	15 (1.1%)	15 (1.1%)	19 (1.4%)
missing (M)	20 (1.5%)	20 (1.5%)	20 (1.5%)
BUSCO eukaryota_odb10			
complete (C)	240 (of 255, 94.1%)	239 (of 255, 93.7%)	239 (of 255, 93.7%)
single copy (S)	240 (94.1%)	239 (93.7%)	239 (93.7%)
duplicated (D)	0 (0.0%)	0 (0.0%)	0 (0.0%)
fragmented (F)	9 (3.5%)	9 (3.5%)	10 (3.9%)
missing (M)	6 (2.4%)	7 (2.8%)	6 (2.4%)

Chapter 2 Figures

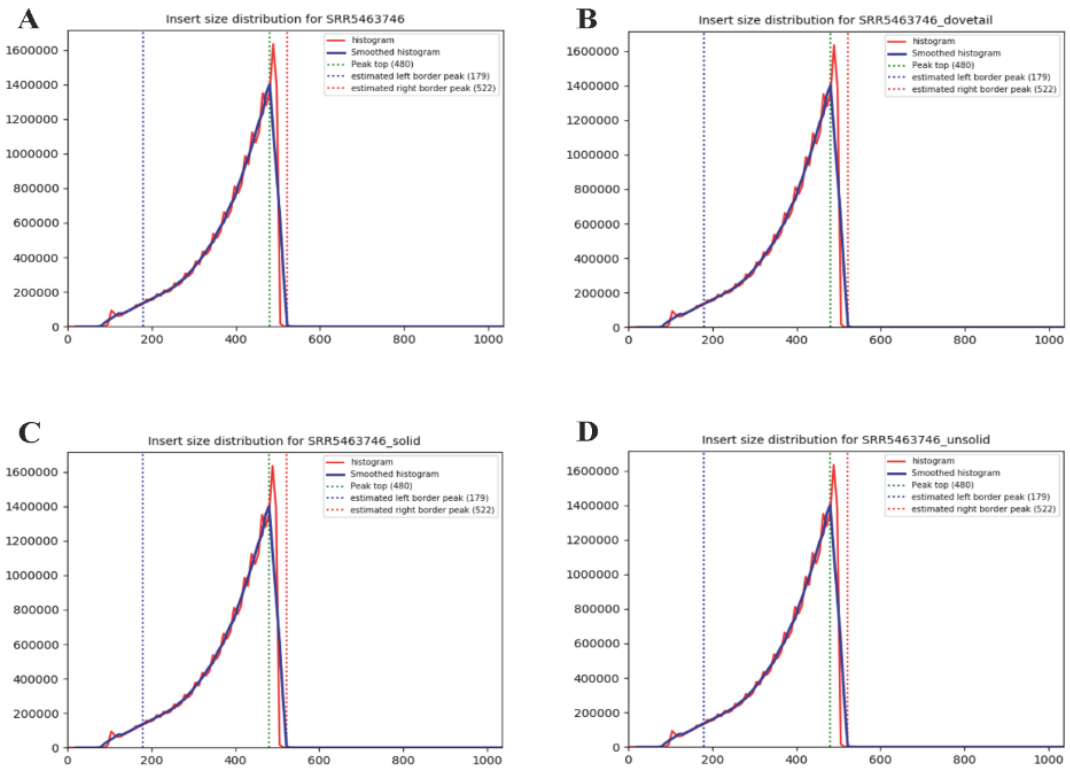


Figure 2.1. Paired end library (SRA accession: SRR5463746) insert size distributions for (A) the published *C. virescens* assembly (GenBank assembly accession: GCA_002382865.1), (B) the Dovetail assembly, (C) *C. virescens_solid* assembly, and (D) *C. virescens_unsolid* assembly.

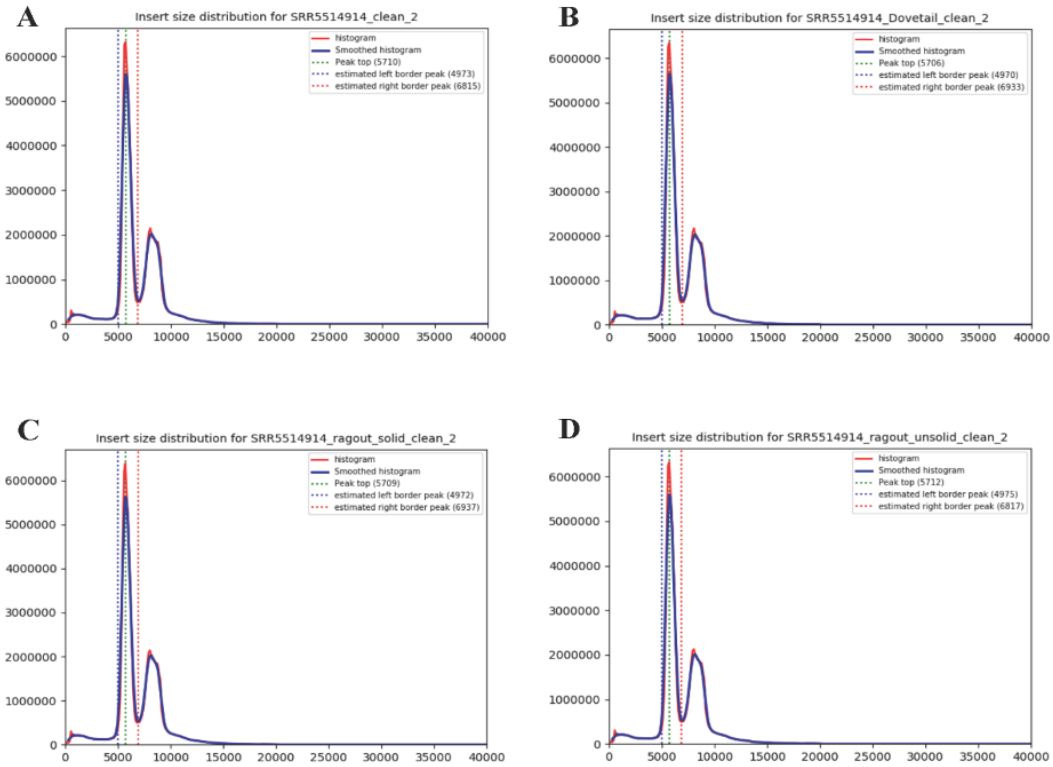


Figure 2.2. Mate pair library (SRA accession: SRR5514914) insert size distributions of for (A) the published *C. virescens* assembly (GenBank assembly accession: GCA_002382865.1), (B) the Dovetail assembly, (C) *C. virescens_solid* assembly, and (D) *C. virescens_unsolid* assembly.

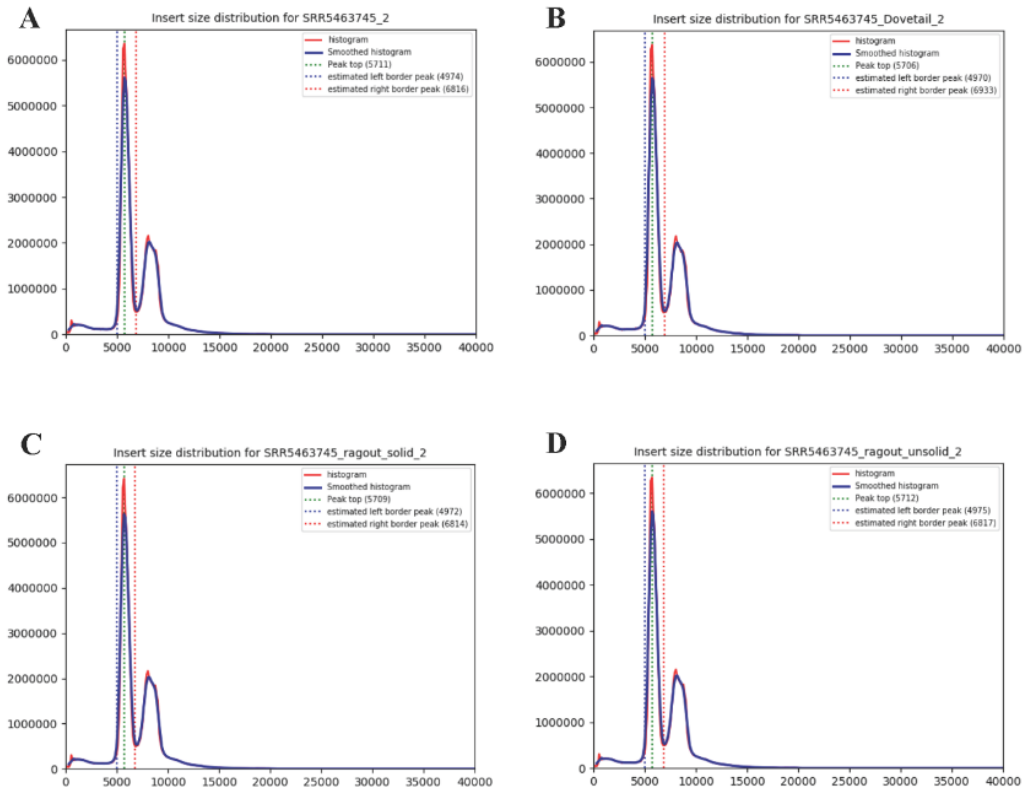


Figure 2.3. Mate pair library (SRA accession: SRR5463745) insert size distributions for (A) the published *C. virescens* assembly (GenBank assembly accession: GCA_002382865.1), (B) the Dovetail assembly, (C) *C. virescens_solid* assembly, and (D) *C. virescens_unsolid* assembly.

Chapter 2 Supplemental Data Files

Data S1. Contiguity of publicly available insect genome assemblies downloaded from NCBI and i5k (Last updated: 2020-07-16).

Data S2. Contiguities and BUSCO scores for publicly available Lepidopteran genome assemblies downloaded from NCBI and i5k (Last updated: 2020-07-16).

Data S3. (A) Information of *C. virescens* linkage markers from Fritz et al, 2016. (B) Mapped positions of linkage map markers (Fritz et al, 2016) in the published *C. virescens* assembly (GenBank assembly accession: GCA_002382865.1). (C) Mapped positions for all linkage map markers (Fritz et al, 2016) in the *C. virescens_solid* assembly. (D) Mapped positions for all linkage map markers (Fritz et al, 2016) in the *C. virescens_unsolid* assembly, which also indicates the start and the end positions of each scaffold in the joining and extension process.

Chapter 4 Tables

Table 4.1. Read counts, mapping rate, mean GC content and mean read alignment depth for the *Chloridea* individual sequenced data.

Individual	Read counts	Aligned reads	Mapping rate	Mean GC%	Mean read alignment depth
Hvtob1	90603390	58219977	64.26%	37.14%	21.81
Hvtob2	120802980	12626191	10.45%	55.50%	4.73
Hvtob3	90079124	55773328	61.92%	37.25%	20.89
Hvtob4	108766444	68382856	62.87%	36.71%	25.61
Hvtob5	86251034	54765443	63.50%	37.05%	20.51
Hvtob6	97733324	62326006	63.77%	36.92%	23.34
Hvtob7	88515734	57610669	65.09%	37.63%	21.58
Hvtob8	76238042	48960032	64.22%	38.22%	18.34
Hvtob9	96508622	61604266	63.83%	36.69%	23.07
Hvtob10	86201656	54388609	63.09%	36.66%	20.37
Hvtom1	82764288	53406571	64.53%	36.74%	20.00
Hvtom2	65680044	42578692	64.83%	37.01%	15.95
Hvtom3	93052412	26132894	28.08%	39.12%	9.79
Hvtom4	71871838	46080688	64.12%	36.26%	17.26
Hvtom5	81418754	51715930	63.52%	36.30%	19.37
Hvtom6	79624504	51197874	64.30%	37.26%	19.18
Hvtom7	74144916	48493095	65.40%	38.60%	18.16
Hvtom8	65304014	41931341	64.21%	37.31%	15.71
Hvtom9	73928488	47312411	64.00%	37.77%	17.72
Hvtom10	76395478	46796728	61.26%	36.40%	17.53
Hstom1	95390518	41383841	43.38%	36.60%	13.40
Hstom2	88264662	38076616	43.14%	37.62%	12.33
Hstom3	84544478	37807385	44.72%	37.73%	12.24
Hstom4	90183650	39016377	43.26%	36.51%	12.64
Hstom5	79743484	36061021	45.22%	37.21%	11.68

Hstom6	76388008	34312672	44.92%	37.10%	11.11
Hstom7	82642622	35326685	42.75%	37.67%	11.44
Hstom8	80413868	36358978	45.21%	36.93%	11.78
Hstom9	72587270	30696257	42.29%	38.80%	9.94
Hstom10	78139142	35448766	45.37%	37.17%	11.48
wildHs1	81189342	36527875	44.99%	36.93%	11.83
wildHs2	102416286	46319610	45.23%	36.66%	15.00
wildHs3	72946856	32460660	44.50%	36.80%	10.51
wildHs4	69928876	31792497	45.46%	36.85%	10.30
wildHs5	90366874	41186844	45.58%	36.90%	13.34
wildHs6	81726938	36966480	45.23%	36.99%	11.97
wildHs7	78406300	36369674	46.39%	36.97%	11.78
wildHs8	86122158	39840134	46.26%	36.92%	12.90
wildHs9	68177566	31064637	45.56%	36.28%	10.06
wildHs10	78255260	35953142	45.94%	36.81%	11.64

Table 4.2. Total number of *Chloridea* larvae, mean infestation rate and sampling time per larva of 8-week field collection in summer 2020.

Crop	Weeks	Total # of larvae	Average infestation rate	Larvae/min
Tobacco	1	0	0.00%	0.00
Tobacco	2	0	0.00%	0.00
Tobacco	3	5	4.00%	0.11
Tobacco	4	7	6.67%	0.20
Tobacco	5	7	9.33%	0.18
Tobacco	6	10	9.33%	0.21
Tobacco	7	9	10.67%	0.16
Tobacco	8	21	20.00%	0.22
Tomatillo	1	25	14.67%	0.21
Tomatillo	2	25	20.00%	0.15
Tomatillo	3	7	8.00%	0.06
Tomatillo	4	11	12.00%	0.06
Tomatillo	5	8	6.67%	0.05
Tomatillo	6	21	21.33%	0.30
Tomatillo	7	39	34.67%	0.23
Tomatillo	8	62	45.33%	0.38
Chickpeas	1	0	0.00%	0.00
Chickpeas	2	1	1.33%	0.02
Chickpeas	3	1	1.33%	0.02
Chickpeas	4	3	4.00%	0.07
Chickpeas	5	7	8.00%	0.10
Chickpeas	6	8	9.33%	0.13
Chickpeas	7	18	22.67%	0.19
Chickpeas	8	14	12.00%	0.23
Cotton	1	0	0.00%	0.00
Cotton	2	0	0.00%	0.00
Cotton	3	0	0.00%	0.00
Cotton	4	0	0.00%	0.00
Cotton	5	0	0.00%	0.00
Cotton	6	1	1.33%	0.01
Cotton	7	0	0.00%	0.00
Cotton	8	1	1.33%	0.01

Table 4.3. Total number of *Chloridea* larvae, mean infestation rate and sampling time per larva of 8-week field collection in summer 2021.

Crop	Weeks	Total # of larvae	Average infestation rate	Larvae/min
Tobacco	1	1	1.33%	0.03
Tobacco	2	3	4.00%	0.08
Tobacco	3	15	18.67%	0.38
Tobacco	4	15	14.67%	0.21
Tobacco	5	7	8.00%	0.17
Tobacco	6	4	5.33%	0.13
Tobacco	7	16	13.33%	0.37
Tobacco	8	12	14.67%	0.63
Tomatillo	1	6	5.33%	0.07
Tomatillo	2	11	9.33%	0.11
Tomatillo	3	10	8.00%	0.11
Tomatillo	4	13	13.33%	0.13
Tomatillo	5	7	6.67%	0.08
Tomatillo	6	14	12.00%	0.17
Tomatillo	7	26	24.00%	0.34
Tomatillo	8	56	44.00%	0.51
Chickpeas	1	2	1.33%	0.06
Chickpeas	2	19	16.00%	0.53
Chickpeas	3	9	9.33%	0.21
Chickpeas	4	18	18.67%	0.38
Chickpeas	5	27	25.33%	0.54
Chickpeas	6	44	37.33%	0.75
Chickpeas	7	85	40.00%	1.10
Chickpeas	8	91	45.33%	1.03
Cotton	1	0	0.00%	0.00
Cotton	2	1	1.33%	0.02
Cotton	3	2	2.67%	0.02
Cotton	4	0	0.00%	0.00
Cotton	5	1	1.33%	0.01
Cotton	6	0	0.00%	0.00
Cotton	7	1	1.33%	0.01
Cotton	8	2	2.67%	0.02

Table 4.4. Collection information of the 20 *C. virescens* and 10 *C. subflexa* trapped males used for WGS. The other 10 wild *C. subflexa* were collected from tomatillo plots and reared to adults in the lab.

Sample ID	**Collection Year**	**Farm**	**Crop**	**Species**
Hv_tob_1	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tob_2	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tob_3	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tob_4	2021	Beltsville	tobacco	*C. virescens*
Hv_tob_5	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tob_6	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tob_7	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tob_8	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tob_9	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tob_10	2021	Upper Marlboro	tobacco	*C. virescens*
Hv_tom_1	2021	Upper Marlboro	tomatillo	*C. virescens*
Hv_tom_2	2021	Upper Marlboro	tomatillo	*C. virescens*
Hv_tom_3	2021	Keedysville	tomatillo	*C. virescens*
Hv_tom_4	2021	Beltsville	tomatillo	*C. virescens*
Hv_tom_5	2021	Upper Marlboro	tomatillo	*C. virescens*
Hv_tom_6	2021	Upper Marlboro	tomatillo	*C. virescens*
Hv_tom_7	2021	Beltsville	tomatillo	*C. virescens*
Hv_tom_8	2021	Upper Marlboro	tomatillo	*C. virescens*
Hv_tom_9	2021	Upper Marlboro	tomatillo	*C. virescens*
Hv_tom_10	2021	Upper Marlboro	tomatillo	*C. virescens*
Hs_tom_1	2021	Keedysville	tomatillo	*C. subflexa*
Hs_tom_2	2020	Keedysville	tomatillo	*C. subflexa*
Hs_tom_3	2020	Keedysville	tomatillo	*C. subflexa*
Hs_tom_4	2021	Keedysville	tomatillo	*C. subflexa*
Hs_tom_5	2020	Keedysville	tomatillo	*C. subflexa*
Hs_tom_6	2021	Beltsville	tomatillo	*C. subflexa*
Hs_tom_7	2020	Keedysville	tomatillo	*C. subflexa*
Hs_tom_8	2020	Keedysville	tomatillo	*C. subflexa*
Hs_tom_9	2020	Keedysville	tomatillo	*C. subflexa*
Hs_tom_10	2020	Keedysville	tomatillo	*C. subflexa*

Table 4.5. Repeat classifications in the genome assembly of *C. subflexa* (GenBank assembly accession: GCA_022398575.1).

Class	Number of elements	Length occupied	Percentage of sequence
SINEs	95	9237 bp	0.00%
ALUs	0	0 bp	0.00%
MIRs	0	0 bp	0.00%
LINEs	9104	912145 bp	0.20%
LINE1	0	0 bp	0.00%
LINE2	2107	256934 bp	0.06%
L3/CR1	86	28953 bp	0.01%
LTR elements	4760	655845 bp	0.14%
ERVL	0	0 bp	0.00%
ERVL-MaLRs	0	0 bp	0.00%
ERV_class I	0	0 bp	0.00%
ERV_class II	0	0 bp	0.00%
DNA elements	2495	188038 bp	0.04%
hAT-Charlie	97	7580 bp	0.00%
TcMar-Tigger	0	0 bp	0.00%
Unclassified	360812	20207494 bp	4.33%
Total interspersed repeats		21972759 bp	4.71%
Small RNA	106	9643 bp	0.00%
Satellites	0	0 bp	0.00%
Simple repeats	56631	2270113 bp	0.49%
Low complexity	8519	386117 bp	0.08%

Table 4.6. Fisher-exact tests for allele frequency changes of nonsynonymous SNPs between *C. virescens* populations, with applying a Bonferroni corrected alpha level (corrected $\alpha = 0.05/36$ compared pairs = 0.00139).

ORs	Amino acid position	Nucleotide position	Allele freq in Hv tobacco	Allele freq in Hv tomatillo	Fisher-exact test p-value
OR6	276	826	61.11% A, 38.89% C	66.67% A, 33.33% C	1
OR6	276	828	16.67% C, 83.33% T	11.11% C, 88.89% T	1
OR13	41	122	27.78% A, 72.22% C	22.22% A, 77.78% C	1
OR13	55	163	33.33% A, 66.67% G	16.67% A, 83.33% G	0.44
OR13	106	316	72.22% A, 27.78% T	55.56% A, 44.44% T	0.48
OR14	49	146	5.56% G, 94.44% C	33.33% G, 66.67% C	0.09
OR14	56	167	66.67% A, 33.33% G	50% A, 50% G	0.50
OR14	57	169	66.67% A, 11.11% G, 22.22% C	44.44% A, 11.11% G, 44.44% C	0.36
OR14	74	220	50% A, 50% G	44.44% A, 55.56% G	1
OR14	82	247	55.56% A, 44.44% G	55.56% A, 44.44% G	1
OR14	123	370	50% G, 50% T	55.56% G, 44.44% T	1
OR14	135	403	88.89% A, 11.11% T	88.89% A, 11.11% T	1
OR14	167	499	55.56% G, 44.44% T	55.56% G, 44.44% T	1
OR14	180	538	16.67% C, 83.33% T	22.22% C, 77.78% T	1
OR14	187	560	94.44% A, 5.56% C	72.22% A, 27.78% C	0.18
OR14	198	592	50% G, 50% T	55.56% A, 44.44% T	1
OR14	207	620	33.33% C, 66.67% T	38.89% C, 61.11% T	1
OR14	212	634	55.56% A, 44.44% G	61.11% A, 38.89% G	1
OR14	229	685	38.89% G, 61.11% T	44.44% G, 55.56% T	1
OR14	264	790	22.22% G, 77.78% T	5.56% G, 94.44% T	0.34
OR14	267	800	22.22% C, 77.78% T	27.78% C, 72.22% T	1
OR14	267	801	5.56% A, 94.44% C	16.67% A, 83.33% C	0.60
OR14	311	931	22.22% A, 77.78% G	11.11% A, 88.89% C	0.66

Table 4.6. Continued.

ORs	Amino acid position	Nucleotide position	Allele freq in Hv tobacco	Allele freq in Hv tomatillo	Fisher-exact test p-value
OR14	314	940	83.33% A, 16.67% G	55.56% A, 44.44% G	0.15
OR14	314	942	61.11% G, 38.89% C	83.33% G, 16.67% C	0.26
OR14	348	1042	72.22% C, 27.78% T	66.67% C, 33.33% T	1
OR14	355	1065	33.33% G, 66.67% C	22.22% G, 77.78% C	0.71
OR14	388	1162	66.67% G, 33.33% T	77.78% G, 22.22% T	0.71
OR16	53	158	72.22% C, 27.78% T	55.56% C, 44.44% T	0.49
OR16	259	775	88.89% A, 11.11% C	83.33% A, 16.67% C	1
OR16	260	779	27.78% A, 72.22% T	16.67% A, 83.33% T	0.70
OR16	260	780	27.78% A, 72.22% G	16.67% A, 83.33% G	0.70
OR16	264	790	66.67% A, 33.33% C	55.56% A, 44.44% C	0.73
OR16	282	844	5.56% A, 94.44% G	22.22% A, 77.78% G	0.34
OR16	307	920	88.89% C, 11.11% T	72.22% C, 27.78% T	0.40
OR16	307	921	16.67% A, 83.33% G	5.56% A, 94.44% G	0.60

Table 4.7. Fisher-exact tests for allele frequency changes of nonsynonymous SNPs between *C. subflexa* populations, with applying a Bonferroni corrected alpha level (corrected $\alpha = 0.05/17$ compared pairs = 0.00294).

ORs	**Amino acid position**	**Nucleotide position**	**Allele freq in Hs from tomatillo traps**	**Allele freq in wild Hs***	**Fisher-exact test p-value**
OR6	208	623	75% C, 25% T	90% C, 10% T	0.41
OR6	208	624	75% A, 25% C	90% A, 10% C	0.41
OR14	41	121	70% A, 30% G	75% A, 25% G	1
OR14	45	133	65% A, 35% G	65% A, 35% G	1
OR14	49	146	35% A, 65% C	35% A, 65% C	1
OR14	49	147	35% A, 65% G	35% A, 65% G	1
OR14	60	179	65% C, 35% T	65% C, 35% T	1
OR14	60	180	65% A, 35% T	65% A, 35% T	1
OR14	116	346	90% C, 10% T	85% C, 15% T	1
OR14	117	349	60% C, 40% T	60% C, 40% T	1
OR14	273	818	30% A, 70% G	30% A, 70% G	1
OR14	326	978	25% G, 75% T	25% G, 75% T	1
OR14	439	1315	75% A, 25% C	75% A, 25% C	1
OR16	9	25	35% A, 65% G	30% A, 70% G	1
OR16	114	340	45% A, 55% G	50% A, 50% G	1
OR16	140	418	65% C, 35% T	60% C, 40% T	1
OR16	371	1111	20% A, 80% T	40% A, 60% T	0.30

***** These wild Hs were collected as larvae from tomatillo plots and reared to adults in the lab.

Table 4.8. Chi-square tests of amino acid variation distribution at an α-level of 0.05. No amino acid variations were detected in *C. subflexa* OR13.

	C. virescens		***C. subflexa***	
	chi-square	p-value	chi-square	p-value
OR6	1.12	0.57	4.13	0.13
OR13	0.9	0.64	NA	NA
OR14	2.76	0.25	3.08	0.21
OR16	3.38	0.18	4.74	0.09

Chapter 4 Figures

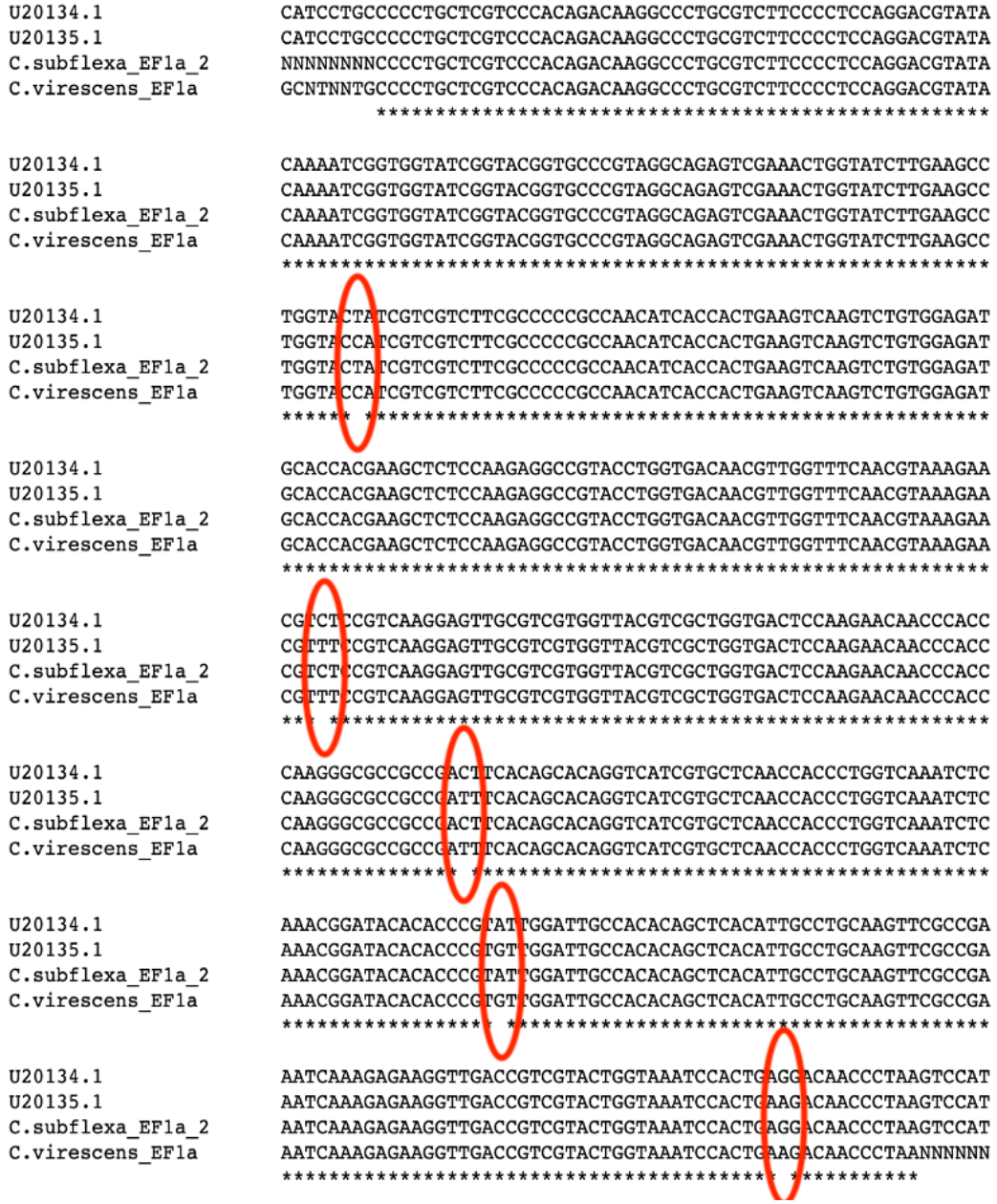
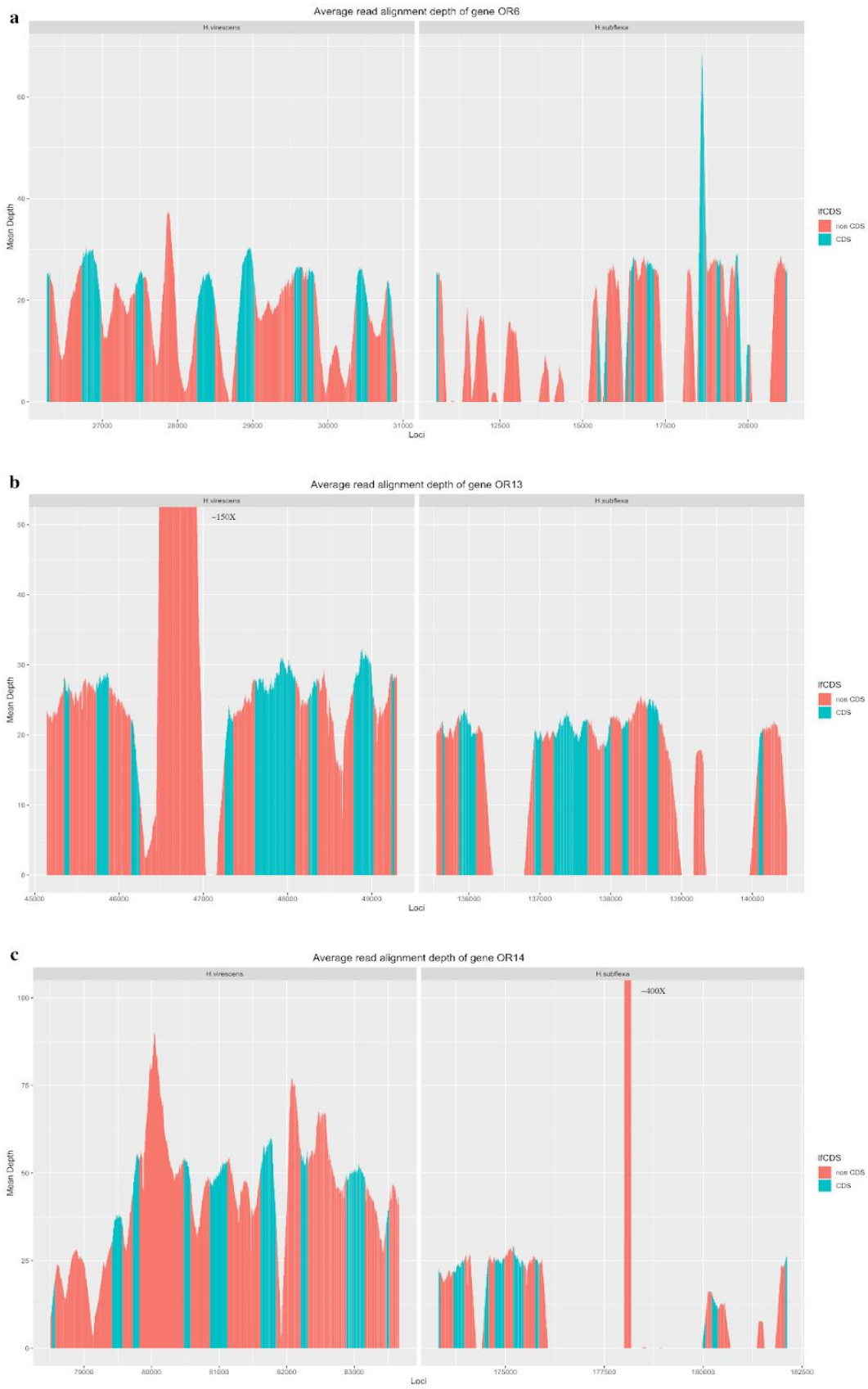


Figure 4.1. Sequence alignment of the region of EF-1 that was used in species identification. U20134.1 and U20135.1 are the nucleotide sequences of EF-1 in *C. subflexa* and *C. virescens* that were obtained from NCBI (Cho et al, 1995). The 5 interspecific variant sites were marked in red circles.


a
Average read alignment depth of gene OR6
H.virescens
H.subflexa
Mean Depth
Loci
lfCDS
non CDS
CDS
b
Average read alignment depth of gene OR13
~150X
c
Average read alignment depth of gene OR14
~400X

d

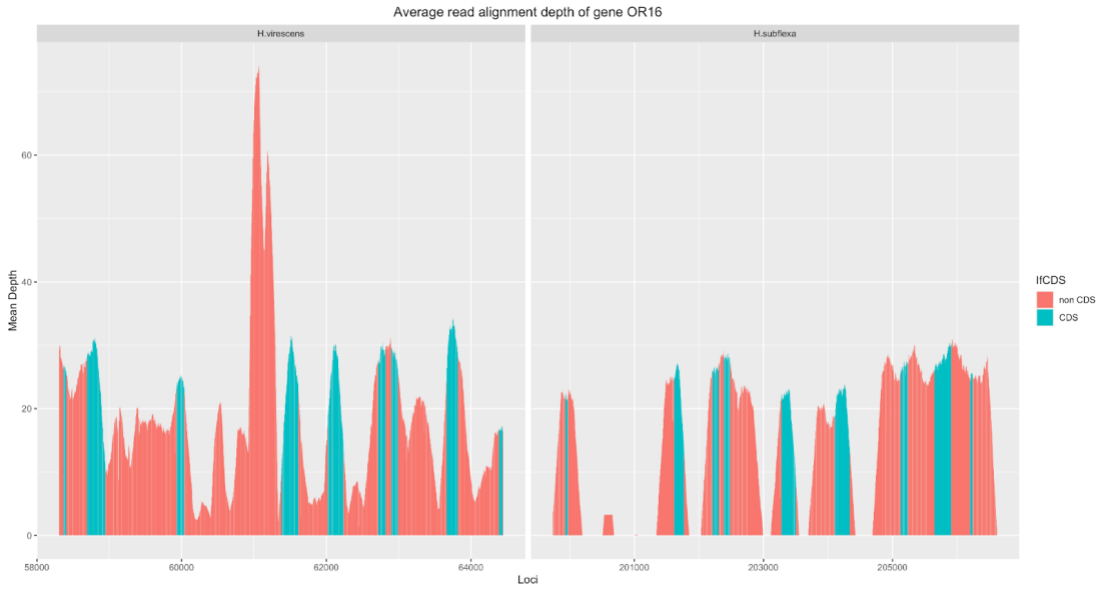


Figure 4.2. Average read alignment depth of genes OR6 (a), OR13 (b), OR14 (c) and OR16 (d) in *C. virescens* and *C. subflexa* populations. Coding regions (CDS) shown as blue. Reads were mapped against the reference genomes of *C. virescens* (GenBank assembly accession: GCA_002382865.1) or *C. subflexa* (GenBank assembly accession: GCA_022398575.1) using Bowtie2, and read depth was measured by samtools depth. Averages were calculated among the 18 individuals of *C. virescens* and the 20 individuals of *C. subflexa*.

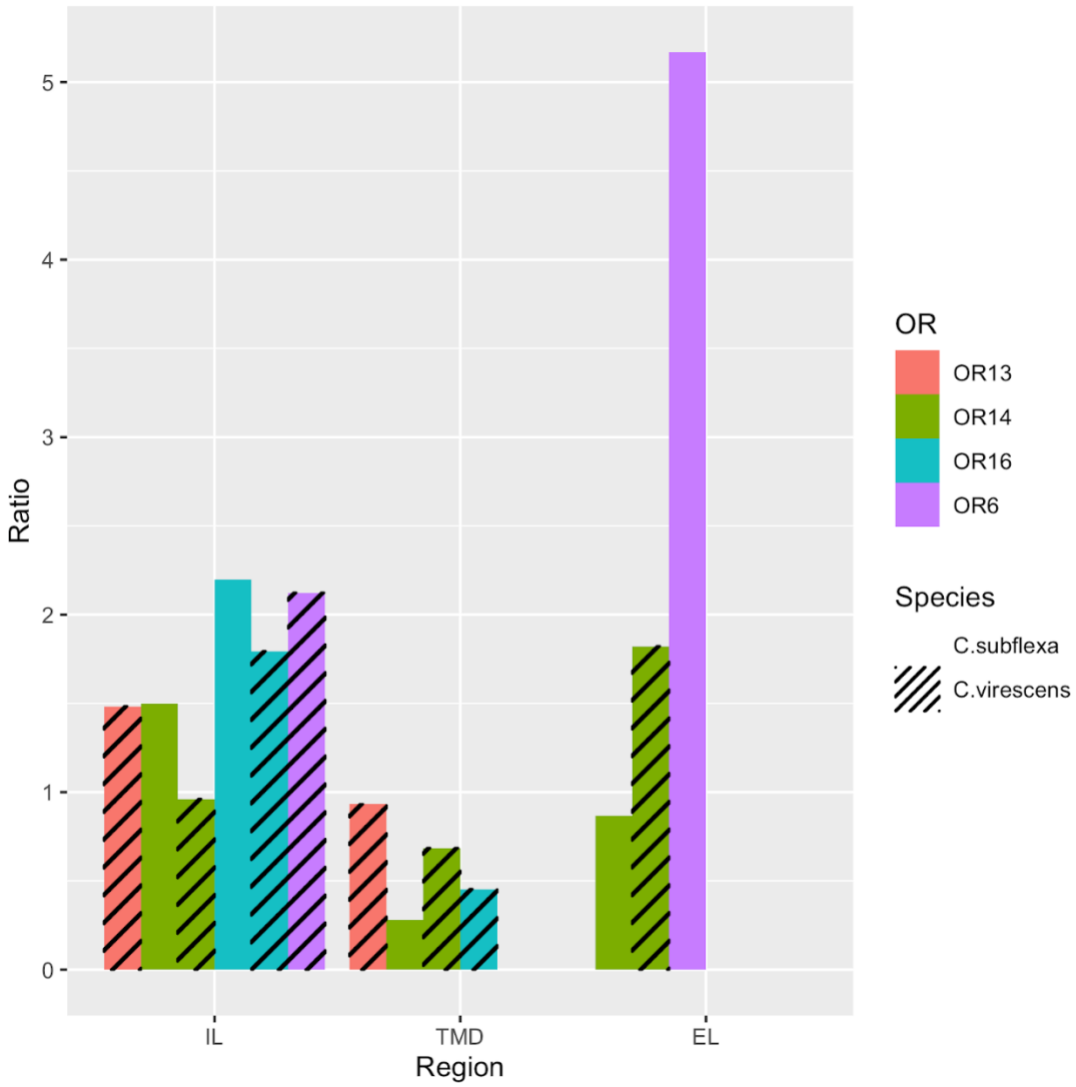


Figure 4.3. Ratio of the relative amino acid variations of OR6, OR13, OR14 and OR16 in *C. virescens* and *C. subflexa* among different regions. No amino acid variations were detected in *C. subflexa* OR13. The ratio of each region was calculated as the number of observed amino acid variations divided the number of expected amino acid variations. Expected amino acid variations: total number of amino acid variations * length of the domain / length of the protein. A ratio of 1 indicates that amino acid changes occurred at the same rate across the entire protein. IL: intracellular loops, TMD: transmembrane domains, EL: extracellular loops.

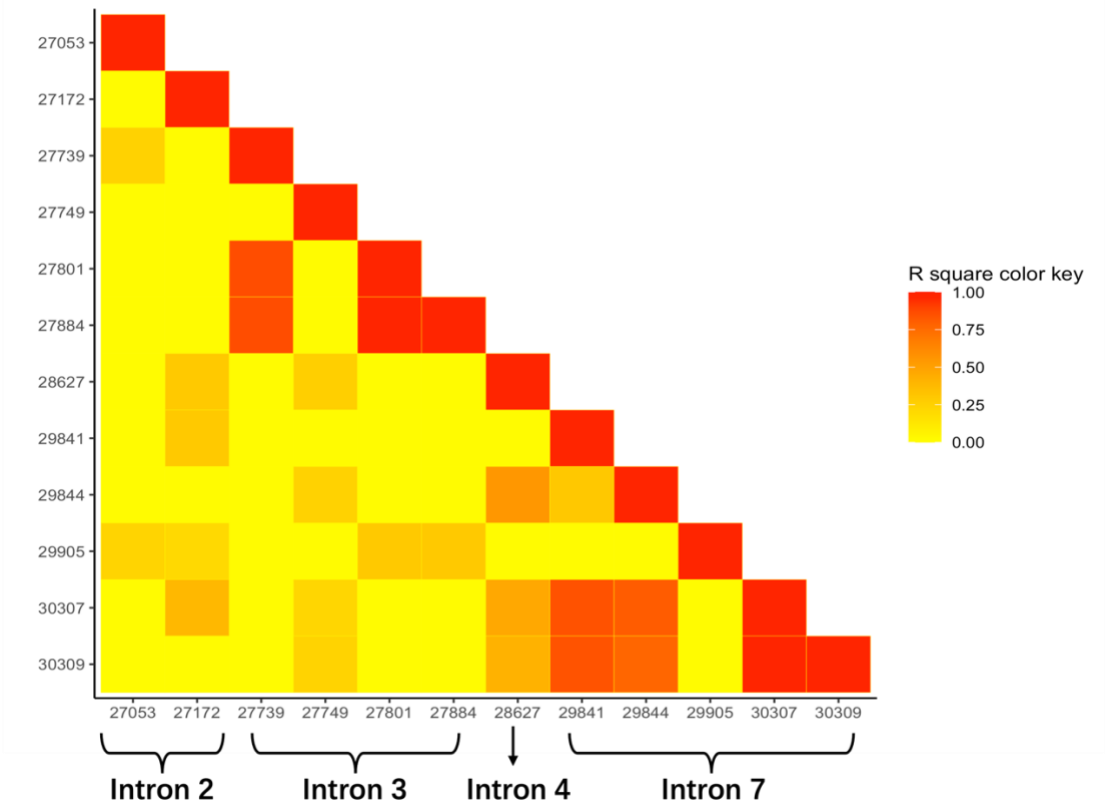


Figure 4.4. Linkage disequilibrium heatmap of top 5% highly divergent SNPs in *C. virescens* OR6.

Chapter 4 Supplemental Data Files

Data S4. Fasta file of Scaffold_4031_rc.Scaffold_489.

Bibliography

Alonge, Michael, et al. "RaGOO: fast and accurate reference-guided scaffolding of draft genomes." *Genome Biology* 20.1 (2019): 1-17.

Baker, Thomas Charles, et al. "A comparison of responses from olfactory receptor neurons of *Heliothis subflexa* and *Heliothis virescens* to components of their sex pheromone." *Journal of Comparative Physiology A* 190.2 (2004): 155-165.

Bao, Ergude, Tao Jiang, and Thomas Girke. "AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references." *Bioinformatics* 30.12 (2014): i319-i328.

Barthel, Andrea, et al. "Immune modulation enables a specialist insect to benefit from antibacterial withanolides in its host plant." *Nature Communications* 7 (2016): 12530.

Benelli, Giovanni, et al. "Sex pheromone aerosol devices for mating disruption: challenges for a brighter future." *Insects* 10.10 (2019): 308.

Bioinformatics, Babraham. "FastQC: a quality control tool for high throughput sequence data." *Cambridge, UK: Babraham Institute* (2011).

Blanco, Carlos A., et al. "Densities of Heliothis virescens and Helicoverpa zea (Lepidoptera: Noctuidae) in three plant hosts." *Florida Entomologist* (2007): 742-750.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30.15 (2014): 2114-2120.

Bushmanova, Elena, et al. "rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data." *GigaScience* 8.9 (2019): giz100.

Cabral-de-Mello, Diogo Cavalcanti, and František Marec. "Universal fluorescence in situ hybridization (FISH) protocol for mapping repetitive DNAs in insects and other arthropods." *Molecular Genetics and Genomics* 296.3 (2021): 513-526.

Callahan, F. E., et al. "High level expression of “male specific” pheromone binding proteins (PBPs) in the antennae of female noctuiid moths." *Insect biochemistry and molecular biology* 30.6 (2000): 507-514.

Cantarel, Brandi L., et al. "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes." *Genome research* 18.1 (2008): 188-196.

Cao, Song, et al. "A single point mutation causes one-way alteration of pheromone receptor function in two Heliothis species." *Iscience* 24.9 (2021): 102981.

Cardé, R. T., et al. "Sex pheromone specificity as a reproductive isolating mechanism among the sibling speciesArchips argyrospilus andA. mortuanus and other sympatric tortricine moths (Lepidoptera: Tortricidae)." *Journal of Chemical Ecology* 3.1 (1977):71-84.

Cardé, Ring T., and K. F. Haynes. "Structure of the pheromone communication." *Advances in insect chemical ecology* 283 (2004).

Cardé, RING T. "Moth navigation along pheromone plumes." *Pheromone communication in moths: evolution, behavior and application. University of California Press, Berkeley, California* (2016): 173-189.

Carraher, Colm, et al. "Towards an understanding of the structural basis for insect olfaction by odorant receptors." *Insect Biochemistry and Molecular Biology* 66 (2015): 31-41.

Chang, Ching-Ho, and Amanda M Larracuente. "Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the *Drosophila melanogaster* Y Chromosome." *Genetics* vol. 211,1 (2019): 333-348. doi:10.1534/genetics.118.301765

Chang, Hetan, et al. "Candidate odorant binding proteins and chemosensory proteins in the larval chemosensory tissues of two closely related noctuidae moths, Helicoverpa armigera and H. assulta." *PLoS one* 12.6 (2017): e0179243.

Chapman, R. F. "Foraging and food choice in phytophagous insects." *Chemical ecology* 1 (2009): 72-101.

Chen, Lihui, et al. "Detecting host-plant volatiles with odorant receptors from Grapholita molesta (Busck)(Lepidoptera: Tortricidae)." *Journal of agricultural and food chemistry* 68.9 (2020): 2711-2717.

Childers, Anna K., et al. "The USDA-ARS Ag100Pest Initiative: High-Quality Genome Assemblies for Agricultural Pest Arthropod Research." *Insects* 12.7 (2021): 626.

Cho, Soowon, et al. "A highly conserved nuclear gene for low-level phylogenetics: elongation factor-1 alpha recovers morphology-based tree for heliothine moths." *Molecular biology and evolution* 12.4 (1995): 650-656.

Cho, Soowon, et al. "Molecular phylogenetics of heliothine moths (Lepidoptera: Noctuidae: Heliothinae), with comments on the evolution of host range and pest status." *Systematic Entomology* 33.4 (2008): 581-594.

Coates, Brad S., Craig A. Abel, and Omaththage P. Perera. "Estimation of long terminal repeat element content in the *Helicoverpa zea* genome from high-throughput sequencing of bacterial artificial chromosome pools." *Genome* 60.4 (2017): 310-324.

Conceição, Inês C., and Montserrat Aguadé. "High incidence of interchromosomal transpositions in the evolutionary history of a subset of Or genes in Drosophila." *Journal of molecular evolution* 66.4 (2008): 325-332

Cunningham, John Paul, and Myron P. Zalucki. "Understanding heliothine (Lepidoptera: Heliothinae) pests: what is a host plant?." *Journal of Economic Entomology* 107.3 (2014): 881-896.

Danecek, Petr, et al. "The variant call format and VCFtools." *Bioinformatics* 27.15 (2011): 2156-2158.

Danecek, Petr, et al. "Twelve years of SAMtools and BCFtools." *Gigascience* 10.2 (2021): giab008.

Davey, John W., et al. "No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions." *Evolution letters* 1.3 (2017): 138-154.

Del Angel, Victoria Dominguez, et al. "Ten steps to get started in Genome Assembly and Annotation." *F1000Research* 7 (2018).

Deng, Jian-Yu, et al. "Enhancement of attraction to sex pheromones of Spodoptera exigua by volatile compounds produced by host plants." *Journal of chemical ecology* 30.10 (2004): 2037-2045.

Dong, Kun, et al. "Key site residues of pheromone-binding protein 1 involved in interacting with sex pheromone components of Helicoverpa armigera." *Scientific reports* 7.1 (2017): 1-9.

Edgar, Robert C. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic acids research* 32.5 (2004): 1792-1797.

Ellis, Emily A., et al. De novo genome assemblies of butterflies. *GigaScience* 10 (2021): 1-8.

Excoffier, Laurent, Peter E. Smouse, and JM1205020 Quattro. "Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data." *Genetics* 131.2 (1992): 479-491.

Eyun, Seong-il, et al. "Evolutionary history of chemosensory-related gene families across the Arthropoda." *Molecular biology and evolution* 34.8 (2017): 1838-1862.

Feder, Jeffrey L., Stewart H. Berlocher, and Susan B. Opp. "Sympatric host-race formation and speciation in Rhagoletis (Diptera: Tephritidae): a tale of two species for Charles D." *Genetic structure and local adaptation in natural insect populations*. Springer, Boston, MA, 1998. 408-441.

Feuk, Lars, et al. "Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies." *PLoS Genetics* 1.4 (2005): e56.

Fritz, Megan L., et al. "Application of a dense genetic map for assessment of genomic responses to selection and inbreeding in *Heliothis virescens*." *Insect Molecular Biology* 25.4 (2016): 385-400.

Fritz, Megan L., et al. "Contemporary evolution of a Lepidopteran species, Heliothis virescens, in response to modern agricultural practices." *Molecular ecology* 27.1 (2018): 167-181.

Forêt, Sylvain, and Ryszard Maleszka. "Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (Apis mellifera)." *Genome research* 16.11 (2006): 1404-1413.

Fu, Yu, et al. "The genome of the Hi5 germ cell line from *Trichoplusia ni*, an agricultural pest and novel model for small RNA biology." *Elife* 7 (2018): e31628.

Gaston, Lyle K., et al. "Controlling the pink bollworm by disrupting sex pheromone communication between adult moths." *Science* 196.4292 (1977): 904-905.

Gemmell, Neil J. "Repetitive DNA: genomic dark matter matters." *Nature Reviews Genetics* 22.6 (2021): 342-342

Gong, Da-Ping, et al. "The odorant binding protein gene family from the genome of silkworm, Bombyx mori." *BMC genomics* 10.1 (2009): 1-14.

Gopinath, G. R., et al. "A hybrid reference-guided de novo assembly approach for generating *Cyclospora* mitochondrion genomes." *Gut Pathogens* 10.1 (2018): 15.

Gould, Fred, et al. "Sexual isolation of male moths explained by a single pheromone response QTL containing four receptor genes." *Proceedings of the National Academy of Sciences* (2010): 200910945.

Grabherr, Manfred G., et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome." *Nature Biotechnology* 29.7 (2011): 644.

Grabherr, Manfred G., et al. "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data." *Nature biotechnology* 29.7 (2011): 644.

Graur, Dan, Amy Katherine Sater, and Tim F. Cooper. *Molecular and genome evolution*. Massachusetts, USA: Sinauer Associates, Incorporated, 2016.

Groot, Astrid T., et al. "Experimental evidence for interspecific directional selection on moth pheromone communication." *Proceedings of the National Academy of Sciences* 103.15 (2006): 5858-5863.

Groot, Astrid T., et al. "Genetic differentiation across North America in the generalist moth Heliothis virescens and the specialist H. subflexa." *Molecular Ecology* 20.13 (2011): 2676-2692.

Groot, Astrid T., et al. "QTL analysis of sex pheromone blend differences between two closely related moths: insights into divergence in biosynthetic pathways." *Insect biochemistry and molecular biology* 39.8 (2009): 568-577.

Groot, Astrid, et al. "Male and female antennal responses in *Heliothis virescens* and *H. subflexa* to conspecific and heterospecific sex pheromone compounds." *Environmental Entomology* 34.2 (2005): 256-263.

Groot, Astrid T., Teun Dekker, and David G. Heckel. "The genetic basis of pheromone evolution in moths." *Annual review of entomology* 61 (2016): 99-117.

Guerrero, Sarahlynne, Julieta Brambila, and Robert L. Meagher. "Efficacies of four pheromone-baited traps in capturing male Helicoverpa (Lepidoptera: Noctuidae) moths in northern Florida." *Florida Entomologist* 97.4 (2014): 1671-1678.

Guo, Hao, et al. "Three pheromone-binding proteins help segregation between two Helicoverpa species utilizing the same pheromone components." *Insect Biochemistry and Molecular Biology* 42.9 (2012): 708-716.

Hekmat-Scafe, Daria S., et al. "Genome-wide analysis of the odorant-binding protein gene family in Drosophila melanogaster." *Genome research* 12.9 (2002): 1357-1369.

Herre, Margaret, et al. "Non-canonical odor coding in the mosquito." *Cell* 185.17 (2022): 3104-3123.

Holt, Robert A., et al. "The genome sequence of the malaria mosquito Anopheles gambiae." *Science* 298.5591 (2002): 129-149.

Huang, Shengfeng, Mingjing Kang, and Anlong Xu. "HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly." *Bioinformatics* 33.16 (2017): 2577-2579.

Hunt, Martin, et al. "REAPR: a universal tool for genome assembly evaluation" *Genome Biology* 14 (2013): R47.
i5K Consortium. "The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment." *Journal of Heredity* 104.5 (2013): 595-600.

Jiang, Xiao-Jing, et al. "Sequence similarity and functional comparisons of pheromone receptor orthologs in two closely related Helicoverpa species." *Insect Biochemistry and Molecular Biology* 48 (2014): 63-74.

Johansson, Björn G., and Therésa M. Jones. "The role of chemical communication in mate choice." *Biological Reviews*82.2 (2007): 265-289.

Johnston, J. Spencer, Angelina Bernardini, and Carl E. Hjelmen. "Genome size estimation and quantitative cytogenetics in insects." *Insect Genomics*. Humana Press, New York, NY, 2019. 15-26.

Jousselin, Emmanuelle, and Marianne Elias. "Testing host-plant driven speciation in phytophagous insects: a phylogenetic perspective." *arXiv preprint arXiv:1910.09510* (2019).

Kamvar, Zhian N., Javier F. Tabima, and Niklaus J. Grünwald. "Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction." *PeerJ* 2 (2014): e281.

Karlson, Peter, and Martin Lüscher. "'Pheromones': a new term for a class of biologically active substances." *Nature*. 183.4653 (1959): 55.

Karpe, Snehal Dilip, Vikas Tiwari, and Sowdhamini Ramanathan. "InsectOR—Webserver for sensitive identification of insect olfactory receptor genes from non-model genomes." *PloS one* 16.1 (2021): e0245324.

Kawano, Shoichi, et al. "Extrafloral nectaries and chemical signals of Fallopia japonica and Fallopia sachalinensis (Polygonaceae), and their roles as defense systems against insect herbivory." *Plant Species Biology* 14.2 (1999): 167-178.

Keightley, Peter D., et al. "Estimation of the spontaneous mutation rate in Heliconius melpomene." *Molecular biology and evolution* 32.1 (2015): 239-243.

Kirkness, Ewen F., et al. "Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle." *Proceedings of the National Academy of Sciences* 107.27 (2010): 12168-12173.

Koehl, M. A. R. "The fluid mechanics of arthropod sniffing in turbulent odor plumes." *Chemical senses* 31.2 (2006): 93-105.

Kolmogorov, Mikhail, et al. "Chromosome assembly of large and complex genomes using multiple references." *Genome Research* 28.11 (2018): 1720-1732.

Kolmogorov, Mikhail, et al. "Ragout—a reference-assisted assembly tool for bacterial genomes." *Bioinformatics* 30.12 (2014): i302-i309.

Korf, Ian. "Gene finding in novel genomes." *BMC bioinformatics* 5.1 (2004): 1-9.

Kozak, Krzysztof M., et al. "Multilocus species trees show the recent adaptive radiation of the mimetic Heliconius butterflies." *Systematic Biology* 64.3 (2015): 505-524.

Krieger, J., et al. "A candidate olfactory receptor subtype highly conserved across different insect orders." *Journal of Comparative Physiology A* 189.7 (2003): 519-526.

Krieger, J., et al. "Genes encoding candidate pheromone receptors in a moth (*Heliothis virescens*)." *Proceedings of the National Academy of Sciences* 101.32 (2004): 11845-11850.

Krogh, Anders, et al. "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." *Journal of molecular biology* 305.3 (2001): 567-580.

Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9.4 (2012): 357-359.

Larter, Nikki K., Jennifer S. Sun, and John R. Carlson. "Organization and function of Drosophila odorant binding proteins." *Elife* 5 (2016): e20242.

Laster, M. L., S. D. Pair, and D. F. Martin. "Acceptance and development of Heliothis subflexa and H. virescens (Lepidoptera: Noctuidae), and their hybrid and backcross progeny on several plant species." *Environmental Entomology* 11.4 (1982): 979-980.

Leal, Walter S. "Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes." *Annu Rev Entomol* 58.1 (2013): 373-391.

Leary, Greg P., et al. "Single mutation to a sex pheromone receptor provides adaptive specificity between closely related moth species." *Proceedings of the National Academy of Sciences* 109.35 (2012): 14081-14086.

Lebreton, Sebastien, et al. "A *Drosophila* female pheromone elicits species-specific long-range attraction via an olfactory channel with dual specificity for sex and food." *BMC biology* 15.1 (2017): 1-14.

Letunic, Ivica, and Peer Bork. "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation." *Nucleic acids research* 49.W1 (2021): W293-W296.

Li, Fei, et al. "Insect genomes: progress and challenges." *Insect Molecular Biology* (2019).

Li, Heng, et al. "The sequence alignment/map format and SAMtools." *Bioinformatics* 25.16 (2009): 2078-2079.

Li, Heng. "Minimap2: pairwise alignment for nucleotide sequences." *Bioinformatics* 34.18 (2018): 3094-3100.

Li, Liang, et al. "Cloning, sequence analysis and spatio-temporal expression of a pheromone binding protein 2 (PBP2) gene from Helicoverpa assulta (Guenée)(Lepidoptera: Noctuidae)." *Acta Entomologica Sinica* 52.11 (2009): 1199-1205.

Light, Douglas M., et al. "Host-plant green-leaf volatiles synergize the synthetic sex pheromones of the corn earworm and codling moth (Lepidoptera)." *Chemoecology* 4.3 (1993): 145-152.

Lischer, Heidi EL, and Kentaro K. Shimizu. "Reference-guided de novo assembly approach improves genome reconstruction for related species." *BMC bioinformatics* 18.1 (2017): 474.

Liu, Binghang, et al. "Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects." *arXiv preprint arXiv:1308.2012* (2013).

Liu, Nai-Yong, et al. "Identification and characterization of three chemosensory receptor families in the cotton bollworm Helicoverpa armigera." *Bmc Genomics* 15.1 (2014): 1-13.

Löfstedt, C., et al. "No linkage between genes controlling female pheromone production and male pheromone response in the European corn borer, Ostrinia nubilalis Hübner (Lepidoptera; Pyralidae)." *Genetics* 123.3 (1989): 553-556.

Löfstedt, Christer. "Moth pheromone genetics and evolution." *Phil. Trans. R. Soc. Lond. B* 340.1292 (1993): 167-177.

Luo, Ruibang, et al. "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." *Gigascience* 1.1 (2012): 18.

Malo, Edi A., et al. "Evaluation of commercial pheromone lures and traps for monitoring male fall armyworm (Lepidoptera: Noctuidae) in the coastal region of Chiapas, Mexico." *Florida Entomologist* (2001): 659-664.

Manoharan, Malini, et al. "Comparative genomics of odorant binding proteins in Anopheles gambiae, Aedes aegypti, and Culex quinquefasciatus." *Genome biology and evolution* 5.1 (2013): 163-180.

Marçais, Guillaume, and Carl Kingsford. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." *Bioinformatics* 27.6 (2011): 764-770.

Marçais, Guillaume, et al. "MUMmer4: a fast and versatile genome alignment system." *PLoS Computational Biology* 14.1 (2018): e1005944.

Martin, Marcel. "Cutadapt removes adapter sequences from high-throughput sequencing reads." *EMBnet. journal* 17.1 (2011): 10-12.

Matthews, Benjamin J., et al. "Improved reference genome of *Aedes aegypti* informs arbovirus vector control." *Nature* 563.7732 (2018): 501-507.

Minkin, Ilia, and Paul Medvedev. "Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ." *Nature communications* 11.1 (2020): 1-11.

Morgulis, Aleksandr, et al. "WindowMasker: window-based masker for sequenced genomes." *Bioinformatics* 22.2 (2005): 134-141.

Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Nei, Masatoshi, and Wen-Hsiung Li. "Mathematical model for studying genetic variation in terms of restriction endonucleases." *Proceedings of the National Academy of Sciences* 76.10 (1979): 5269-5273.

Nozawa, Masafumi, and Masatoshi Nei. "Evolutionary dynamics of olfactory receptor genes in Drosophila species." *Proceedings of the National Academy of Sciences* 104.17 (2007): 7122-7127.

Ochieng, S., K. Park, and T. Baker. "Host plant volatiles synergize responses of sex pheromone-specific olfactory receptor neurons in male *Helicoverpa zea*." *Journal of Comparative Physiology A* 188.4 (2002): 325-333.

Omasits, Ulrich, et al. "Protter: interactive protein feature visualization and integration with experimental proteomic data." *Bioinformatics* 30.6 (2014): 884-886.

Oppenheim, Sara J., Fred Gould, and Keith R. Hopper. "The genetic architecture of a complex ecological trait: host plant use in the specialist moth, *Heliothis subflexa*." *Evolution: International Journal of Organic Evolution* 66.11 (2012): 3336-3351.

Oppenheim, Sara J., Fred Gould, and Keith R. Hopper. "The genetic architecture of ecological adaptation: intraspecific variation in host plant use by the lepidopteran crop pest Chloridea virescens." *Heredity* 120.3 (2018): 234-250.

Pages, H., et al. "Package 'Biostrings'." *Bioconductor* (2013): 18129

Pearce, Stephen L., et al. "Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and invasive *Helicoverpa* pest species." *BMC Biology* 15.1 (2017): 63.

Pelosi, Paolo, and Rosario Maida. "Odorant-binding proteins in insects." *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 111.3 (1995): 503-514.

Pelosi, Paolo, et al. "Beyond chemoreception: diverse tasks of soluble olfactory proteins in insects." *Biological Reviews* 93.1 (2018): 184-200.

PERCY-CUNNINGHAM, JEAN E., and J. A. MacDonald. "Biology and Ultrastructure of Sex Pheromone–Producing Glands." *Pheromone biochemistry*. 1987. 27-75.

Pfeifer, Bastian, et al. "PopGenome: an efficient Swiss army knife for population genomic analyses in R." *Molecular biology and evolution* 31.7 (2014): 1929-1936.

Phelan, P. L. "Evolution of mate-signaling in moths: phylogenetic considerations and predictions from the asymmetric tracking hypothesis." *The evolution of mating systems in insects and arachnids (1997)*: 240-256.

Phelan, P. L. "Evolution of sex pheromones and the role of asymmetric tracking." *Insect chemical ecology: an evolutionary approach* (1992): 265-314.

Poelchau, Monica, et al. "The i5k Workspace@ NAL—enabling genomic data access, visualization and curation of arthropod genomes." *Nucleic acids research* 43.D1 (2014): D714-D719.

Powell, Jerry A. "Biological and taxonomic studies in tortricine moths, with reference to species in California." *Biological and taxonomic studies in tortricine moths, with reference to species in California.* 32 (1964).

Pringle, Elizabeth G., et al. "Synteny and chromosome evolution in the Lepidoptera: evidence from mapping in *Heliconius melpomene*." *Genetics* 177.1 (2007): 417-426.

Purcell, Shaun, et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *The American journal of human genetics* 81.3 (2007): 559-575.

Rabieh, M. M. "Biodiversity of noctuid moths (Lepidoptera: Noctuidae) in the agroecosystems of Mashhad County." *Biodiversity Int J* 2.2 (2018): 147-151.

Regnier, Fred E., and John H. Law. "Insect pheromones." *Journal of Lipid Research* 9.5 (1968): 541-551.

Richards, Stephen, and Shwetha C. Murali. "Best practices in insect genome sequencing: what works and what doesn't." *Current Opinion in Insect Science* 7 (2015): 1-7.

Robertson, Hugh M., Coral G. Warr, and John R. Carlson. "Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster." *Proceedings of the National Academy of Sciences* 100.suppl_2 (2003): 14537-14542.

Roelofs, Wendell L., and Andre Comeau. "Sex pheromone specificity: taxonomic and evolutionary aspects in Lepidoptera." *Science* 165.3891 (1969): 398-400.

Roelofs, Wendell, et al. "Sex pheromone production and perception in European corn borer moths is determined by both autosomal and sex-linked genes." *Proceedings of the National Academy of Sciences* 84.21 (1987): 7585-7589.

Roelofs, Wendell L., and Alejandro P. Rooney. "Molecular genetics and evolution of pheromone biosynthesis in Lepidoptera." *Proceedings of the National Academy of Sciences* 100.16 (2003): 9179-9184.

Rothschild, G. H. L. "Control of oriental fruit moth (Cydia molesta (Busck)(Lepidoptera, Tortricidae)) with synthetic female pheromone." *Bulletin of Entomological Research* 65.3 (1975): 473-490.

Rouyar, Angéla, et al. "Unexpected plant odor responses in a moth pheromone system." *Frontiers in physiology* 6 (2015): 148.

Sakurai, Takeshi, Shigehiro Namiki, and Ryohei Kanzaki. "Molecular and neural mechanisms of sex pheromone reception and processing in the silkmoth Bombyx mori." *Frontiers in physiology* 5 (2014): 125.

Salzberg, Steven L., et al. "GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* 22 (2012): 557-567.

Sánchez-Gracia, A., F. G. Vieira, and J. Rozas. "Molecular evolution of the major chemosensory gene families in insects." *Heredity* 103.3 (2009): 208-216.

Schneeberger, Korbinian, et al. "Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes." *Proceedings of the National Academy of Sciences* 108.25 (2011): 10249-10254.

Seixas, Fernando A., Nathaniel B. Edelman, and James Mallet. "Synteny-based genome assembly for 16 species of *Heliconius* butterflies, and an assessment of structural variation across the genus." *Genome Biology and Evolution*, 2021.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. "BUSCO: assessing genome assembly and annotation completeness." *Gene Prediction*. Humana, New York, NY, 2019. 227-245.

Sheck, A. L., and F. Gould. "The genetic basis of host range in Heliothis virescens: larval survival and growth." *Entomologia Experimentalis et Applicata* 69.2 (1993): 157-172.

Sinnwell, Jason P., Daniel J. Schaid, and Maintainer Jason P. Sinnwell. "Package 'haplo. stats'." (2022).

Sitchawat, Tawatchai, and Thurston Richard. "Heliothis species on tobacco cultivars and Physalis species in Kentucky." *Annals of the Entomological Society of America* 73.4 (1980): 375-377.

Sivashankari, Selvarajan, and Piramanayagam Shanmughavel. "Comparative genomics-a perspective." *Bioinformation* 1.9 (2007): 376.

Slone, Jesse D., et al. "Functional characterization of odorant receptors in the ponerine ant, Harpegnathos saltator." *Proceedings of the National Academy of Sciences* 114.32 (2017): 8586-8591.

Smart, Renee, et al. "Drosophila odorant receptors are novel seven transmembrane domain proteins that can signal independently of heterotrimeric G proteins." *Insect biochemistry and molecular biology* 38.8 (2008): 770-780.

Smadja, C., and R. K. Butlin. "On the scent of speciation: the chemosensory system and its role in premating isolation." *Heredity* 102.1 (2009): 77-97.

Smit A. F. A., Hubley, R., and Green, P. (2013). *RepeatMasker Open-4.0.*

Smith, Christopher D., et al. "Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile)." *Proceedings of the National Academy of Sciences* 108.14 (2011): 5673-5678.

Sonenshine, D. E. "Pheromones: Function and Use in Insect and Tick Control." (2016).

Staab, Paul R., and Dirk Metzler. "Coala: an R framework for coalescent simulation." *Bioinformatics* 32.12 (2016): 1903-1904.

Stanke, Mario, et al. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." *Bioinformatics* 24.5 (2008): 637-644.

Sullivan, Susan L., et al. "The chromosomal distribution of mouse odorant receptor genes." *Proceedings of the National Academy of Sciences* 93.2 (1996): 884-888.

Sun, Ya-Lan, et al. "Expression in antennae and reproductive organs suggests a dual role of an odorant-binding protein in two sibling Helicoverpa species." *PloS one* 7.1 (2012): e30040.

Suyama, Mikita, David Torrents, and Peer Bork. "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments." *Nucleic acids research* 34.suppl_2 (2006): W609-W612.

Svensson, Mats. "Sexual selection in moths: the role of chemical communication." *Biological Reviews* 71.1 (1996): 113-135.

Tajima, Fumio. "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." *Genetics* 123.3 (1989): 585-595.

Tamura, Koichiro, Glen Stecher, and Sudhir Kumar. "MEGA11: molecular evolutionary genetics analysis version 11." *Molecular biology and evolution* 38.7 (2021): 3022-3027.

Tanaka, Kana, et al. "Highly selective tuning of a silkworm olfactory receptor to a key mulberry leaf volatile." *Current Biology* 19.11 (2009): 881-890.

Taylor, Katherine L., et al. "Genome evolution in an agricultural pest following adoption of transgenic crops." *Proceedings of the National Academy of Sciences* 118.52 (2021): e2020853118.

Teal, Peter EA, and Andrea Oostendorp. "Effect of interspecific hybridization between Heliothis virescens and H. subflexa (Lepidoptera: Noctuidae) on sex pheromone production by females." *Journal of insect physiology* 41.6 (1995): 519-525.

Thompson, Julie D., Desmond G. Higgins, and Toby J. Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic acids research* 22.22 (1994): 4673-4680.

Thompson, Julie D., Toby J. Gibson, and Des G. Higgins. "Multiple sequence alignment using ClustalW and ClustalX." *Current protocols in bioinformatics* 1 (2003): 2-3.

Thrash, Adam, et al. “Toward a more holistic method of genome assembly assessment.” *BMC Bioinformatics* 21 (2020): 249.

Triant, Deborah A., Scott D. Cinel, and Akito Y. Kawahara. "Lepidoptera genomes: current knowledge, gaps and future directions." *Current Opinion in Insect Science* 25 (2018): 99-105.

Tristram D. Wyatt. *Pheromones and animal behavior: communication by smell and taste*. Cambridge university press, 2003.

Tsitoura, Panagiota, et al. "Expression and membrane topology of Anopheles gambiae odorant receptors in lepidopteran insect cells." *PloS one* 5.11 (2010): e15428.

Vásquez, G. M., et al. "Differential expression of odorant receptor genes involved in the sexual isolation of two Heliothis moths." *Insect molecular biology* 20.1 (2011): 115-124.

Vieira, Filipe G., Alejandro Sánchez-Gracia, and Julio Rozas. "Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution." *Genome biology* 8.11 (2007): 1-16.

Vieira, Filipe G., and Julio Rozas. "Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system." *Genome biology and evolution* 3 (2011): 476-490.

Vogt, Richard G., and Lynn M. Riddiford. "Pheromone binding and inactivation by moth antennae." *Nature* 293.5828 (1981): 161.

Vogt, Richard G., et al. "A comparative study of odorant binding protein genes: differential expression of the PBP1-GOBP2 gene cluster in Manduca sexta (Lepidoptera) and the organization of OBP genes in Drosophila melanogaster (Diptera)." *Journal of Experimental Biology* 205.6 (2002): 719-744.

Vosshall, Leslie B., et al. "A spatial map of olfactory receptor expression in the Drosophila antenna." *Cell* 96.5 (1999): 725-736.

Vosshall, Leslie B., and Bill S. Hansson. "A unified nomenclature system for the insect olfactory coreceptor." *Chemical senses* 36.6 (2011): 497-498.

Vurture, Gregory W., et al. "GenomeScope: fast reference-free genome profiling from short reads." *Bioinformatics* 33.14 (2017): 2202-2204.

Wan, Fanghao, et al. "A chromosome-level genome assembly of Cydia pomonella provides insights into chemical ecology and insecticide resistance." *Nature Communications* 10.1 (2019): 1-14.

Wang, Bing, Yang Liu, and Gui-Rong Wang. "Proceeding from in vivo functions of pheromone receptors: peripheral-coding perception of pheromones from three closely related species, Helicoverpa armigera, H. assulta, and Heliothis virescens." *Frontiers in physiology* 9 (2018): 1188.

Wang, G., et al. "Functional characterization of pheromone receptors in the tobacco budworm Heliothis virescens." *Insect molecular biology* 20.1 (2011): 125-133.

Waterhouse, Robert M., et al. "Evolutionary superscaffolding and chromosome anchoring to improve *Anopheles* genome assemblies." *BMC Biology* 18.1 (2020): 1-20.

Weir, Bruce S., and C. Clark Cockerham. "Estimating F-statistics for the analysis of population structure." *evolution* (1984): 1358-1370.

Wickham, Hadley. "Data analysis." *ggplot2*. Springer, Cham, 2016. 189-201.

Wilson, Edward O. "Chemical Communication in the Social Insects: Insect societies are organized principally by complex systems of chemical signals." *Science* 149.3688 (1965): 1064-1071.

Witzgall, Peter, et al. "Codling moth management and chemical ecology." *Annu. Rev. Entomol.* 53 (2008): 503-522.

Wu, Han, et al. "An odorant receptor and glomerulus responding to farnesene in Helicoverpa assulta (Lepidoptera: Noctuidae)." *Insect biochemistry and molecular biology* 115 (2019): 103106.

Xia, Qingyou, et al. "A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*)." *Science* 306.5703 (2004): 1937-1940.

Xia, Qingyou, et al. "Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*)." *Science* 326.5951 (2009): 433-436.

Xu, Wei, et al. "Chemosensory receptor genes in the Oriental tobacco budworm Helicoverpa assulta." *Insect Molecular Biology* 24.2 (2015): 253-263.

Xue, Wei, et al. "L_RNA_scaffolder: scaffolding genomes with transcripts." *BMC Genomics* 14.1 (2013): 604.

Yang, Ke, et al. "Two single-point mutations shift the ligand selectivity of a pheromone receptor between two closely related moth species." *Elife* 6 (2017): e29100.

Yang, Zhihua, Marie Bengtsson, and Peter Witzgall. "Host plant volatiles synergize response to sex pheromone in codling moth, *Cydia pomonella*." *Journal of chemical ecology* 30.3 (2004): 619-629.

Yang, Ziheng, and Joseph P. Bielawski. "Statistical methods for detecting molecular adaptation." Trends in ecology & evolution 15.12 (2000): 496-503.

Yang, Ziheng. "PAML 4: phylogenetic analysis by maximum likelihood." *Molecular biology and evolution* 24.8 (2007): 1586-1591.

Young, Janet M., and Barbara J. Trask. "The sense of smell: genomics of vertebrate odorant receptors." *Human molecular genetics* 11.10 (2002): 1153-1160.

Zhang, Hui-Jie, et al. "A phylogenomics approach to characterizing sensory neuron membrane proteins (SNMPs) in Lepidoptera." *Insect Biochemistry and Molecular Biology* 118 (2020): 103313.

Zhang, Jin, William B. Walker, and Guirong Wang. "Pheromone reception in moths: from molecules to behaviors." *Progress in Molecular Biology and Translational Science* 130 (2015): 109-128.

Zhang, Shen, et al. "Genome size reversely correlates with host plant range in *Helicoverpa* species." *Frontiers in Physiology* 10 (2019): 29.

Zhang, Tian-Tao, et al. "Characterization of three pheromone-binding proteins (PBPs) of Helicoverpa armigera (Hübner) and their binding properties." *Journal of insect physiology* 58.7 (2012): 941-948.

Zhang, Zhi-Qiang. *Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness*. Magnolia press, 2011.

Zhou, Jing-Jiang, et al. "“Plus-C” odorant-binding protein genes in two Drosophila species and the malaria mosquito Anopheles gambiae." *Gene* 327.1 (2004): 117-129.

Zhou, Jing, et al. "Identification of host-plant volatiles and characterization of two novel general odorant-binding proteins from the legume pod borer, Maruca vitrata Fabricius (Lepidoptera: Crambidae)." *PLoS one* 10.10 (2015): e0141208.