University of Alberta

Phylogenomics of the *Choristoneura fumiferana* species complex (Lepidoptera: Tortricidae)

by

Heather Michelle Bird

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

in

Systematics and Evolution

Department of Biological Sciences

Fall 2013

Edmonton, Alberta

For my parents Art and Dixie Bird,

thank you for your wise advice, understanding,

unconditional love, and support.

Abstract

The phylogenetic relationships of the destructive spruce budworm group of forest pests (*Choristoneura fumiferana* species complex) have previously been explored using allozymes, microsatellites, mitochondrial genes and a nuclear gene, but remain poorly resolved with conflicting topologies. I used mass sampling of single nucleotide polymorphisms (SNPs) across their genome in a genotyping-by-sequencing approach. Over 100 000 SNPs with greater than 75% coverage in 99 specimens (ApeKI restriction enzyme digest) or 144 specimens (PstI-MspI digest) resolved *C. fumiferana*, *C. carnana*, *C. retiniana*, and *C. pinus* as strongly-supported monophyletic species. The most distinct species, *C. pinus*, yielded 945 autapomorphic SNPs, and was definitively placed as basal to the whole species complex, contrasting with previous mtDNA results. The functions of genes homologous to the sequence surrounding the diagnostic *C. pinus* SNPs included detoxification, morphological differences, flight, and sensory perception, providing insights into the genetic basis of species differences.

Acknowledgements

Many people were integral to the inspiration and completion of this research. The person most central to its completion and my growth as a graduate student was my supervisor Felix Sperling. I am grateful for his intellectual insight, enduring patience, and kindness in dealing with me, especially because our different personalities made me a perplexing student for him. A more enthusiastic and supportive supervisor I could not have hoped for.

My deepest thanks are to my family. My parents, Art and Dixie Bird, my brother Phil and his wife Violet and their wee one Noelle who was born during this degree, all of my grandparents, Ted and Shelia Burger, Charlie and Ann Bird, and my dear friends Emily Moss, Brian Leibel, and Andrea Cochrane who we remember fondly.

Special thanks to the members of my supervisory committee, Jocelyn Hall and Paul Stothard, as well as our collaborators, Michel Cusson, Roger Levesque, Brian Boyle, Jérôme Laroche, and especially Lisa Lumley. Without their inspiration and advice, this research could not have taken place. The professors who inspired enthusiasm for grad school while I was an undergraduate, Theresa Burg, Stewart Rood, and Alice Hontela, also have my thanks.

I would like to thank the many people who collected samples for this study including the Sperling lab members, Maya Evenden, Jan Volney, James Weber, Jerry Powell, Chenisse and Colton and Cali Squire, Brian Leibel, people associated with the CFS (Canadian Forest Service), GLFC (Great Lakes Forestry Centre insect production services), FIDS (Forest Insect and Disease Survey a Canadian organization that formed in 1961 and disbanded in 1996), and AESRD (Alberta Environment and Sustainable Resource Development) especially Anina Hundsdörfer.

All of the Sperlingites and larger entomology group at the University of Alberta have my thanks for their continuous help and support. They are the friendliest group I have had the pleasure of working among. Special thanks to Marla Schwarzfeld and Jason Dombroskie who took in a newbie student and helped her feel at home, Sarah Leo, Julian Dupuis, Bryan Brunet, Jasmine Janes, Giovanny Fagua, Marcelo and Karina Brandão, Christianne McDonald, and Christi Jaeger for amusing conversations, laughs, sharing their experiences, and entomological enthusiasm!

The field work and sequencing of this study was funded by the ACA Grants in Biodiversity (supported by the Alberta Conservation Association) and Alberta Innovates – Bio Solutions. A Discovery Grant from the Natural Sciences and Engineering Research Council of Canada to Felix Sperling funded my summer stipend for the first year and the early stages of my research.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

aa: amino acid

bp: basepair

COI: cytochrome c oxidase I

COII: cytochrome c oxidase II

GBS: genotyping by sequencing

GO: gene ontology

GTR: general time-reversible

ML: maximum likelihood

MP: maximum parsimony

mtDNA: mitochondrial DNA

NCBI: National Centre for Biotechnology Information

NJ: neighbour joining

NGS: next generation sequencing

rDNA: ribosomal DNA

SNP: single nucleotide polymorphism

SBW: spruce budworm

UV or MV light: ultraviolet or mercury vapor light

Chapter 1

General Introduction

Taxonomy, the classification and naming of living things in the world around us (Wheeler, 2008), is one of the oldest preoccupations of mankind (Yoon, 2010). Today, species are defined by morphology, life history, mating behaviour, ecology, geographic distribution, and genetic divergence (Mayr, 1982; Wheeler and Meier, 2000; Van Valen, 1976; Andersson, 1990). Accurately curating and delimiting species enables us to monitor biodiversity (Kim and Byrne, 2006), manage conserved habitats and species (Franklin, 1993; Mace, 2004), respond appropriately to invasive or native pests (Strauss et al., 2006), and simply to name the life in our backyards and enjoy nature (Louv, 2008). Species are the foundational unit of analysis in ecology, biogeography, and phylogenetics, but delimitating cryptic species (Rissler and Apodaca, 2007), or closely related species (Brown, 1959), can be challenging when based solely on morphology. In these cases an iterative approach involving multiple lines of evidences (De Queiroz, 2011; Tan et al., 2010), and/or multiple genetic markers is necessary. This is especially true for the spruce budworm (SBW) complex, *Choristoneura fumiferana* (Clemens 1865) (Lumley and Sperling, 2010).

Spruce budworm species are native defoliators of conifers in North America that cause serious economic loss in forestry (Volney and Fleming, 2007). Given roughly one million hectares infected by SBW across Canada in an average year (AESRD, 2013) and, in Alberta, a loss of $420/ha to the government and $1206/ha to a forestry company every 25 years if there is no spraying (ASRD, 2002), the SBW would cause an average economic loss of about $65 million to Canada per year if no management steps were taken. Occurrence and length of outbreaks depend on many factors including species of budworm, host species, and size of host stand (Weber and Schweingruber, 1995; Régnière and Nealis, 2007; Royama et al., 2005; McCullough, 2000), so it is important to distinguish species correctly and understand their interactions in a particular region to be able

to implement appropriate management plans. However, distinguishing these species is difficult because they are morphologically variable and this variation overlaps between species (Freeman, 1967; Powell and De Benedictus, 1995). Ecology and geographical range are useful for identifying SBW species (Lumley and Sperling, 2011), however molecular markers are required to understand their evolutionary history and relationships. Previous studies have used the mitochondrial cytochrome c oxidase I (COI) and II (COII) genes (Sperling and Hickey, 1994; Sperling and Hickey, 1995; Lumley and Sperling, 2011), microsatellites (Lumley et al., 2009; Lumley and Sperling, 2011), and allozymes (Stock and Castrovillo, 1981; Harvey, 1984; Harvey, 1996) to delimit species but produced poorly resolved and contradictory topologies. In cases like this, single genes can produce a misleading topology due to homoplasy, which occurs when two species appear to be sister taxa because of convergence, parallelism, or reversal instead of common descent (Gatesy et al., 2007; Pashley and Ke, 1992). Similarly, introgression of genes, chromosomal segments, or mitochondrial genomes can result in paraphyly, polyphyly, and topologies that misrepresent the true evolutionary history (Bossu and Near, 2009). Horizontal transfer, due to retrotransposons copying and inserting segments of DNA in different parts of the same genome or into another individual's genome (Walsh et al., 2012), is a third reason why more than one gene should be sequenced to produce a species phylogeny that accurately presents the evolutionary history of the organism (Dupuis et al., 2012; Maddison and Ober, 2011; Corl and Ellegren, 2013). Similarily, the sequence data should be analysed with multiple methodologies because different phylogenetic methods can produce conflicting topologies (Huelsenbeck, 1995). By using a genome-wide suite of single nucleotide polymorphisms (SNPs) this study aims to shed more light on species limits and their phylogenetic relationships in this closely related species complex.

1.1. The *Choristoneura fumiferana* complex

1.1.1. Natural History

The geographical distributions of SBW species tend to follow the ranges of their host plants in North America, and many are sympatric in part of their ranges. In the east and across boreal North America, *C. fumiferana* feeds on white (*Picea glauca* (Moench) Voss), black (*Picea mariana* (Miller) Britton, Sterns & Poggenburg), and red spruce (*Picea rubens* Sargent), and balsam fir (*Abies balsamea* (Linnaeus) Miller), *C. pinus pinus* Freeman 1953 feeds on jack pine (*Pinus banksiana* Lambert), and *C. pinus maritima* Freeman 1967 feeds on Virginia (*Pinus virginiana* Miller) and pitch pine (*Pinus rigida* Miller) (Freeman and Stehr, 1967; Harvey, 1984; Volney and Fleming, 2007; Lumley and Sperling, 2011). In the west, there are six species: *C. orae* Freeman 1967 feeds on Sitka spruce (*Picea sitchensis* (Bongard) Carrière) and Pacific silver fir (*A. amabilis* Douglas ex Forbes) along coastal British Columbia and in Alaska; *C. biennis* Freeman 1967 feeds on Engelmann spruce (*Picea engelmannii* Parry ex Engelmann) and subalpine fir (*A. lasiocarpa* (Hooker) Nuttall) in the mountainous regions of British Columbia and Alberta; *C. occidentalis* Freeman 1967 feeds on Douglas fir (*Pseudotsuga menziesii* (Mirbel) Franco) in Alberta and British Columbia down through the southern Rocky Mountains; *C. lambertiana* (Busck 1915) and its subspecies, *C. l. ponderosana* and *C. l. subretiniana*, feed on pines including lodgepole (*Pinus contorta* Douglas), ponderosa (*Pinus ponderosa* Douglas ex Lawson), and sugar pine (*Pinus lambertiana* Douglas) along the southern Rocky Mountains; *C. retiniana* (Walsingham 1879) feeds on white (*A. concolor* (Gordon) Lindley ex Hildebrand) and grand fir (*A. grandis* (Douglas ex Don) Lindley) in northern California and neighbouring states; and finally *C. carnana* (Barnes and Busck 1920) and its subspecies feed on Douglas fir and white fir in California (Freeman and Stehr, 1967; Harvey, 1984; Volney and Fleming, 2007; Lumley and Sperling, 2011).

Species of the spruce budworm complex have a univoltine life cycle. The exceptions are *C. biennis* (Freeman, 1967) and sometimes *C. orae* (Lumley,

2010), which are semivoltine and undergo a second diapause at the end of the third instar, subsequently moulting and emerging as fourth instar larvae in the second spring (Nealis, 2005). Larvae emerge in April when temperatures exceed roughly 15°C, are phototactic and move to the tips of the branches where the spruce and fir feeders mine current-year needle buds (Stark and Borden, 1965; Nealis, 2008). Pine feeders mine the staminate flowers and cones, although some needle mining also occurs (Stark and Borden, 1965). In later feeding stages, larvae spin a silk shelter inside the needle buds causing them to retain their bud caps (Nealis, 2008). After four to six weeks, the larvae reach their final instar, the number of which varies between five and eight, and pupate (Nealis, 2008). Adults eclose approximately two weeks later (Stark and Borden, 1965). *Choristoneura fumiferana, C. orae,* and *C. p. maritima* tend to emerge two weeks earlier in the season than *C. occidentalis* and *C. p. pinus*, and the last to emerge are *C. biennis* and *C. lambertiana* (Freeman and Stehr, 1967; Stark and Borden, 1965). Adults fly from late June through August, mate and deposit eggs, often within 24-36 hours of eclosion (Stark and Borden, 1965). Females oviposit on host plant foliage in multiple masses of 15 to 60 eggs in a scale-like pattern (Nealis, 2008; Lumley, 2010). The eggs hatch within two weeks and the first-instar larvae, without feeding, migrate from needles to protected locations on branches (Nealis, 2008; Stark and Borden, 1965). They spin hibernacula to enter diapause and overwinter, emerging as second instar larvae the following spring (Nealis, 2008; Stark and Borden, 1965).

Morphologically, SBW moths have a reddish-brown or grey mottled forewing, with a wingspread of 18 to 31 mm where the female is slightly larger than the male (Obraztsov, 1962; Freeman, 1967). The western species *C. biennis* and *C. occidentalis* are slightly larger than *C. fumiferana* and the other species, and *C. pinus* is typically one of the smallest (Obraztsov, 1962; Freeman, 1967). In colour, *C. fumiferana* and *C. biennis* tend to be grey to reddish-brown with *C. biennis* darker; *C. occidentalis* are reddish-brown with black scales; *C. pinus* and *C. carnana* are reddish-brown; *C. orae* are reddish-brown to grey; *C. lambertiana* are reddish-brown to tan; and *C. retiniana* are pale goldish-tan or

tawny (Lumley and Sperling, 2011; Freeman, 1967; Powell, 1995). There are some diagnostic characteristics on the aedeagus (Dang, 1985), species differences in female calling time in the evening (Sanders et al., 1977) and different mating pheromone ratios, however the pheromone components overlap between some species (Silk and Kuenen, 1988).

1.1.2. Taxonomic History

The taxonomic history of the SBW complex is complicated. It was initially described as a single species, *Tortrix fumiferana,* in 1865 by Clemens, and was subsequently split into light and dark variants (*fumiferana* and *nigridia* respectively) in 1869 by Robinson (Freeman, 1953). These taxa were merged again and generically reassigned to *Archips fumiferana* in 1929 by Graham, and then *Cacoecia fumiferana* in 1943 by Brown and MacKay, and finally to *Choristoneura fumiferana* in 1947 by Freeman (Freeman, 1947). Six years later Freeman (1953) published a formal description for the pine-feeding form and a redescription for the spruce-balsam form (*C. pinus* and *C. fumiferana* respectively). Soon after this, Obraztsov (1962) moved *C. retiniana* to *Choristoneura*, the species having originally been in *Lozotaenia* then *Archips* then *Cacoecia*. Obraztsov (1962) also added new subspecies to *C. lambertiana* including *C. l. ponderosana*, and *C. l. lindseyana*, although *C. l. lindseyana* was later synonymized with *C. retiniana* (Brown et al., 2005). Originally *C. lambertiana* was placed in *Tortrix*, moved to *Cacoecia*, then back to *Tortrix*, then *Archips*, and finally moved to *Choristoneura* (Obraztsov, 1962). Obraztsov (1962) also described the new species *C. subretiniana* which was later designated a subspecies of *C. lambertiana* (Powell, 1964; Brown et al., 2005). Five years later, Freeman (1967) described five new species, *C. biennis, C. occidentalis, C. orae, C. viridus* (later synonymized with *retiniana*; Powell, 1995), and *C. pinus maritima*, distinguishing it from *C. pinus pinus*.

Currently only eight species (*C. biennis*, *C. carnana*, *C. fumiferana*, *C. occidentalis*, *C. orae*, *C. lambertiana*, *C. pinus*, and *C. retiniana*) are formally recognized in the coniferophagous SBW complex (Brown et al., 2005), although

there is some disagreement about the names of some species (Dombroskie, 2011). In the genus *Choristoneura*, seventeen species are Nearctic (Wang and Yang, 2008), and the eight in the SBW complex are the only ones that feed on conifers. Thirty-eight species are recognized in the genus world-wide (Brown et al., 2005).

1.2. Molecular techniques

The study of systematics is the ordering and rationalizing of relationships among organisms (Hennig, 1966), with taxonomy being the counterpart that focuses on giving consistent names to organisms or groups of them. Together they are used to order the chaos of organisms, extant and extinct, and to illuminate the processes of divergence and convergence that underlie the natural organisation of life. The earliest sketches of this process were derived from morphological characteristics, which are still used today with great success (Zeuner, 1943; MacLeod, 2008; De Meulemeester et al., 2012). However, in cases of morphological convergence or cryptic species, morphological characteristics can be misleading and molecular characteristics may be better for delimiting species (Hedges and Sibley, 1994; Van Oppen et al., 2001; Simmons and Scheffer, 2004; Witt and Hebert, 2000).

One of the earliest molecular techniques used in systematics was isozyme electrophoresis, also known as allozyme electrophoresis. Isozymes are functionally similar forms of an enzyme coded by genes at different loci, or by different alleles at the same locus (allozymes). When proteins are extracted from organisms, electrophoresis can separate and characterize the variants (Murphy et al., 1996), and the extent to which variants are shared is used to calculate evolutionary similarity (Stock and Castrovillo, 1981; Harvey, 1984; Harvey, 1996). However, homoplasy, differences in gene expression, and difficulties in laboratory preparation may result in phylogenies that do not accurately describe the succession of species divergences (Murphy, 1988).

Researchers soon developed methods of characterizing DNA, specifically restriction site analysis, Sanger sequencing, and microsatellites. Restriction site

analysis was performed by digesting specimen DNA with restriction enzymes and hybridizing it to a labeled portion of the genome such as mtDNA (Sperling and Harrison, 1994; Sperling, 1993). In this way the restriction sites were mapped and variation in presence/absence of sites could be characterized for species. Years later, Sanger sequencing became practical and the nucleotide sequence of full genes could be characterized. Sanger sequencing is a chain-termination or dideoxy sequencing method (Sanger et al. 1977) that became automated and highly accurate. The discovery of the polymerase chain reaction (PCR) (Mullis, 1990) allowed similar genes in different species to be easily amplified and sequenced, finally making DNA sequencing a routine part of systematics. The nucleotide sequence of a gene can characterize a larger number of variable sites than previous methods, but is also susceptible to homoplasy (Gatesy et al., 2007; Pashley and Ke, 1992), introgression (Bossu and Near, 2009), and horizontal transfer (Walsh et al., 2012). In many cases, more than one gene is required to produce a true species phylogeny (Dupuis et al., 2012; Maddison and Ober, 2011; Corl and Ellegren, 2013). Some species complexes require even more genotyped variable sites to accurately determine their evolutionary relationships.

One method of increasing sampling across the genome is by characterizing microsatellites. Microsatellites are short tandem repeats, or simple sequence repeats, of mono-, di-, tri-, and tetranucleotide motifs (*i.e.* CACACACACA) that are ubiquitous in eukaryotic genomes (Ellegren, 2004). They mutate frequently and length polymorphisms can be genotyped in multiple individuals (Ellegren, 2004). The high number of alleles that can occur at each locus make microsatellites an efficient way to delimit species and populations (Beacham et al., 2008); however discovering microsatellites and developing primers can be time consuming and challenging (Sinama et al., 2011; Tay et al., 2010).

Within the last ten years, several next generation sequencing technologies have been developed, including Roche/454 (parallelized pyrosequencing), Illumina/Solexa, Life Technologies/SOLiD, Life Technologies/Ion Torrent, and Heliscope BioSciences (Metzker, 2010). We used Illumina sequencing in our study. Illumina sequencing attaches DNA fragments to a glass slide and forms

clusters of the fragment using bridge amplification. Bridge amplification occurs when an attached DNA fragment curls over and attaches to an adjacent primer and is replicated (Metzker, 2010). The sequence of each cluster is detected when fluorophore-tagged reversible terminators are cleaved off each newly incorporated nucleotide during a series of PCR wash cycles (Metzker, 2010). The four nucleotides are labelled with different colour fluorophores so they can be washed through the eight-laned slides and detected simultaneously. Although the length sequenced is shorter (100-250 bp for Illumina HiSeq) than traditional Sanger sequencing (up to1400 bp), the new methods output millions of sequences per run and are more cost and time efficient per bp sequenced (Long et al., 2013; Metzker, 2010). From the massive amount of sequence data produced, hundreds of thousands of variable sites (single nucleotide polymorphisms or SNPs) can be genotyped. Finally, this multitude of markers may accurately resolve the true phylogeny of closely related species, revealing the evolutionary history of the entire genomes (Edwards, 2009).

Once homologous SNPs have been genotyped in all species, diagnostic characters can be determined. A diagnostic locus has a genotype that occurs in all specimens of one species, and occurs in no specimens of any other species (*e.g.* all *C. pinus* have an "A" at that locus, and no other species have A, M, R, or W). When such diagnostic SNPs are mapped on a phylogeny, they may be interpreted as autapomorphies if they represent the derived character state. Synapomorphies, genotypes diagnostic for a clade of more than one species, can similarly be determined. Using the *C. fumiferana* reference genome, we can BLAST search the sequence flanking these apomorphic loci against the NCBI database of protein sequences. If the sequences containing these apomorphic SNPs are homologous to known genes, then we can infer their biological functions. Knowing the functions of these genes deepens our understanding of the potential selective and functional factors that may have led to the fixation of these genetically determined biological innovations, ultimately leading to speciation.

1.3. Thesis overview

Understanding the genetic processes promoting speciation, and the genes that enable sub-populations to fill different ecological niches for than a parent population, is basic to our understanding of the mechanisms of evolution. Pursuit of this understanding has led to the work described in this thesis. Chapter 2 addresses the phylogeny of six SBW species and the use of genotyping-by-sequencing to characterize SNP variation in DNA fragments associated with particular types of restriction sites. This study used data for 102 specimens with DNA fragments associated with ApeKI restriction enzyme sites and 144 specimens associated with the PstI-MspI enzyme pair. It also addresses the advantages and challenges of large SNP datasets and restriction enzyme associated DNA in quantifying divergences and reconstructing phylogenies. Chapter 3 describes the diagnostic SNPs found for each of six SBW species, and the genes associated with the 945 SNPs autapomorphic for *C. pinus*. It also describes how these genes may have potentially been involved in the speciation of *C. pinus,* and how apomorphic SNPs fit into the larger context of species concepts.

The SBW complex is a significant forest pest, motiving substantial research of its evolutionary relationships. However, the close genetic relationships of these species have meant that after many years of research their phylogeny has remained poorly resolved. With this project, we aimed to gain the best estimate of the true evolutionary history of these species that is currently possible without sequencing the entire genome. This research will deepen our understanding of the evolutionary relationships between these species and the biological innovations that may have enabled the process of their species divergence.

In summary, the goals of this thesis were to reassess the phylogenetic relationships among SBW species, evaluate the use of new sequencing technologies, determine the amount of characters needed to delimit these species, and discover genes potentially involved in divergence of SBW species, with a focus on *C. pinus*.

1.4. References

AESRD. 2013. Spruce budworm information page. Available from http://cfs.nrcan.gc.ca/pages/50. Accessed 2013-09-17.

Andersson, L. 1990. The driving force: species concepts and ecology. Taxon. 39, 375–382.

ASRD, 2002. Integrated Spruce Budworm Management Strategy. Available from http://srd.alberta.ca/LandsForests/ForestHealth/ForestPests/CommonTreeInsectsDiseases/documents/integratedsbmstrategy.pdf. Accessed 2013-09-17.

Beacham, T.D., Wetklo, M., Wallace, C., Olsen, J.B., Flannery, B.G., Wenburg, J.K., Templin, W.D., Antonovich, A., Seeb, L.W. 2008. The application of microsatellites for stock identification of Yukon River Chinook salmon. N. Am. J. Fish. Manage. 28, 283-295.

Bossu, C.M., Near, T.J. 2009. Gene trees reveal repeated instances of mitochondrial DNA introgression in orangethroat darters (Percidae: Etheostoma). Syst. Biol. 58, 114-129.

Brown, W.J. 1959. Taxonomic problems with closely related species. Ann. Rev. Entomol. 4, 77-98.

Brown, J.W., Baixeras, J., Brown, R., Horak, M., Komai, F., Metzler, E., Razowski, J., Tuck, K. 2005. World Catalogue of Insects, Volume 5, Tortricidae (Lepidoptera). Apollo Books, Stenstrup, Denmark.

Corl, A., Ellegren, H. 2013. Sampling strategies for species trees: The effects on phylogenetic inference of the number of genes, number of individuals, and whether loci are mitochondrial, sex-linked, or autosomal. Mol. Phylogenet. Evol. 67, 358-366.

Dang, P.T. 1985. Key to adult males of conifer-feeding species on *Choristoneura* Lederer (Lepidoptera: Tortricidae) in Canada and Alaska. Can. Entomol. 117, 1-5.

De Meulemeester, T., Michez, D., Aytekin, A.M., Danforth, B.N. 2012. Taxonomic affinity of halictid bee fossils (Hymenoptera: Anthophila)

based on geometric morphometrics analyses of wing shape. J. Syst. Paleontol. 10, 755-764.

De Queiroz, K. 2011. Branches in the lines of descent: Charles Darwin and the evolution of the species concept. Biol. J. Linn. Soc. 103, 19-35.

Dombroskie, J. 2011. Aspects of Archipine Evolution (Lepidoptera: Tortricidae). PhD dissertation. University of Alberta, Edmonton, Alberta, Canada.

Dupuis, J.R., Roe, A.D., Sperling, F.A.H. 2012. Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. Mol. Ecol. 21, 4422-4436.

Edwards, S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63, 1-19.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. 5, 435-445.

Franklin, J.F. 1993. Preserving biodiversity: species, ecosystems, or landscapes? Ecol. Appl. 3, 202–205.

Freeman, T.N. 1947. The external anatomy of the spruce budworm *Choristoneura fumiferana* (Clem.) (Lepidoptera: Tortricidae). Can. Entomol. 79, 21-31.

Freeman, T.N. 1953. The spruce budworm, *Choristoneura fumiferana* (Clem.) and an allied new species on pine (Lepidoptera: Tortricidae). Can. Entomol. 85, 121-127.

Freeman, T.N. 1967. On coniferophagous species of *Choristoneura* (Lepidoptera: Tortricidae) in North America I. Some new forms on the *Choristoneura* allied to *C. fumiferana*. Can. Entomol. 99, 449-455.

Freeman, T.N., Stehr, G.W. 1967. On coniferophagous species of *Choristoneura* (Lepidoptera: Tortricidae) in North America VI. A summary of the preceding five papers. Can. Entomol. 99, 504-506.

Gatesy, J., DeSalle, R., Wahlberg, N. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst. Biol. 56. 355-363.

Harvey, G.T. 1984. The taxonomy of the coniferophagous *Choristoneura* (Lepidoptera: Tortricidae): a review. In: Sanders, C.J., Stark, R.W.,

Mullins, E.J., Murphy, J. (Eds.) Bangor Recent Advances in Spruce Budworm Research. Proceedings CANUSA Spruce budworms Research Symposium., Maine, USA. Canadian Forestry Service, Headquarters, Ottawa, Canada, pp. 16-48.

Harvey, G.T. 1996. Genetic relationships among *Choristoneura* species (Lepidoptera: Tortricidae) in North America as revealed by isozyme studies. Can. Entomol. 128, 245-262.

Hedges, S.B., Sibley, C.G. 1994. Molecules vs. morphology in avian evolution: the case of the "pelecaniform" birds. Proc. Natl. Acad. Sci. 91, 9861-9865.

Hennig, W. 1966. Phylogenetic Systematics. University of Illinois Press, Urbana, IL, USA.

Huelsenbeck, J.P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44, 17-48.

Kim, K.C., Byrne, L.B. 2006. Biodiversity loss and the taxonomic bottleneck: emerging biodiversity science. Ecol. Res. 21, 794-810.

Long, S.W., Williams, D., Valson, C., Cantu, C.C., Cernoch, P., Musser, J.M., Olsen, R.J. 2013. A genomic day in the life of a clinical microbiology laboratory. J. Clin. Microbiol. 51, 1272-1277.

Louv, R. 2008. Last Child in the Woods: Saving Our Children From Nature-Deficit Disorder. Algonquin Books, Chapel Hill, NC, USA.

Lumley, L.M., 2010. Species delimitation in the *Choristoneura fumiferana* species complex (Lepidoptera: Tortricidae). Ph.D. dissertation. University of Alberta, Edmonton, Alberta, Canada.

Lumley, L.M., Davis, C.S., Sperling, F.A.H. 2009. Isolation and characterization of eight microsatellite loci in the spruce budworm species *Choristoneura fumiferana* and *Choristoneura occidentalis*, and cross-species amplification in related tortricid moths. Conservation Genet. Resour. 1, 501-504.

Lumley, L.M. Sperling, F.A.H. 2010. Integrating morphology and mitochondrial DNA for species delimitation within the spruce budworm (*Choristoneura*

fumiferana) cryptic species complex (Lepidoptera: Tortricidae). Syst. Entomol. 35, 416-428.

Lumley, L.M., Sperling, F.A.H. 2011. Utility of microsatellites and mitochondrial DNA for species delimitation in the spruce budworm (*Choristoneura fumiferana*) species complex (Lepidoptera: Tortricidae). Mol. Phylogenet. Evol. 58, 232-243.

Mace, G.M. 2004. The role of taxonomy in species conservation. Phil. Trans. R. Soc. Lond. B. 359, 711-719.

MacLeod, N. 2008. Understanding morphology in systematic contexts, three-dimensional specimen ordination and recognition. In: Wheeler, Q.D. (Ed.) The New Taxonomy. CRC Press, NY, USA, pp.143-209.

Maddison, D.R., Ober, K.A. 2011. Phylogeny of minute carabid beetles and their relatives based upon DNA sequence data (Coleoptera, Carabidae, Trechitae). Zookeys. 147, 229-260.

Mayr, E. 1982. The Growth of Biological Thought: Diversity, Evolution, and Inheritance. Belknap Press, Cambridge, MA, USA.

McCullough, D.G. 2000. A review of factors affecting the population dynamics of jack pine budworm (*Choristoneura pinus pinus* Freeman). Popul. Ecol. 42, 243-256.

Metzker, M.L. 2010. Sequencing technologies – the next generation. Nat. Rev. Genet. 11, 31-46.

Mullis, K.B. 1990. The unusual origin of the polymerase chain reaction. Sci. Am. 262. 56-61, 64-65.

Murphy, R.W. 1988. The problematic phylogenetic analysis of interlocus heteropolymer isozyme characters: a case study from sea snakes and cobras. Can. J. Zool. 66, 2628-2633.

Murphy, R.W., Sites, J.W., Buth, D.G., Haufler, C.H. 1996. Proteins: Isozyme electrophoresis. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.) Molecular Systematics, Second Edition. Sinauer Associates, Inc., Sunderland, MA, USA, pp. 51-120.

Nealis, V.G. 2005. Diapause and voltinism in western and 2-year-cycle spruce budworms (Lepidoptera: Tortricidae) and they hybrid progeny. Can. Entomol. 137, 584-597.

Nealis, V.G. 2008. Spruce budworms, *Choristoneura* Lederer (Lepidoptera: Tortricidae). In: Capinera, J. (Ed.) Encyclopedia of Entomology, second ed. Springer, Dordrecht, pp. 3524-3531.

Obraztsov, N.S. 1962. New species and subspecies of North American Archipini, with notes on other species (Lepidoptera, Tortricidae). Am. Mus. Novit. 2101.

Pashley, D.P., Ke, L.D. 1992. Sequence evolution in mitochondrial ribosomal and ND-1 genes in Lepidoptera: implications for phylogenetic analyses. Mol. Biol. Evol. 9, 1061-1075.

Powell, J.A. 1964. Univ. Calif. Publ. Entomol. 32, 181.

Powell, J.A. 1995. Biosystematic Studies of Conifer-feeding *Choristoneura* (Lepidoptera: Tortricidae) in the Western United States. University of California Press, Berkeley, CA, USA.

Powell, J.A., De Benedictus, J.A. 1995. Evolutionary interpretation, taxonomy and nomenclature. In: Powell, J.A. (Ed.), Biosystematic Studies of Conifer-feeding *Choristoneura* (Lepidoptera: Tortricidae) in the Western United States. University of California Press, Berkeley, CA, USA, pp.219–275.

Régnière, J., Nealis, V.G. 2007. Ecological mechanisms of population change during outbreaks of the spruce budworm. Ecol. Entomol. 32, 461-477.

Rissler, L.J., Apodaca, J.J. 2007. Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). Syst. Biol. 56, 924-942.

Royama, T. MacKinnon, W.E., Kettela, E.G., Carter, N.E., Hartling, L.K. 2005. Analysis of spruce budworm outbreak cycles in New Brunswick, Canada, since 1952. Ecology. 86, 1212-1224.

Sanders, C.J., Daterman, G.E., Ennis, T.J. 1977. Sex pheromone responses of *Choristoneura* spp. and their hybrids (Lepidoptera: Tortricidae). Can. Entomol. 109, 1203-1220.

Sanger, F., Nicklen, S., Coulson, A.R. 1977. DNA sequencing with chain terminating inhibitors. Proc. Nat. Acad. Sci. USA. 74, 5463-5467.

Silk, P.J., Kuenen, L.P.S. 1988. Sex pheromones and behavioral biology of the coniferophagous *Choristoneura*. Ann. Rev. Entomol. 33, 83-101.

Simmons, R.B., Scheffer, S.J. 2004. Evidence of cryptic species within the pest *Copitarsia decolora* (Guenée) (Lepidoptera: Noctuidae). Ann. Entomol. Soc. Am. 97, 675-680.

Sinama, M., Dubut, V., Costedoat, C., Gilles, A., Junker, M., Malause, T., Martin, J-M, Nève, G., Pech, N., Schmitt, T., Zimmermann, M., Meglécz, E. 2011. Challenges of microsatellite development in Lepidoptera: *Euphydryas aurinia* (Nymphalidae) as a case study. Eur. J. Entomol. 108, 261-266.

Sperling, F.A.H. 1993. Mitochondrial DNA variation and Haldane's rule in the *Papilio glaucus* and *P. troilus* species group. Heredity. 71, 227-233.

Sperling, F.A.H., Harrison, R.G. 1994. Mitochondrial DNA variation within and between species of the *Papilio machaon* group of swallowtail butterflies. Evolution. 48, 408-422.

Sperling, F.A.H., Hickey, D.A. 1994. Mitochondrial DNA sequence variation in the spruce budworm species complex (*Choristoneura*: Lepidoptera) Mol. Biol. Evol. 11, 656-665.

Sperling, F.A.H., Hickey, D.A. 1995. Amplified mitochondrial DNA as a diagnostic marker for species of conifer-feeding *Choristoneura* (Lepidoptera: Tortricidae). Can. Entomol. 127, 277-288.

Stark, R.W., Borden, J.H. 1965. Life history of *Choristoneura lambertiana subretiniana* Obraztsov (Lepidoptera: Tortricidae) attacking lodgepole pine. Can. Entomol. 97, 684-690.

Strauss, S.Y., Webb, C.O., Salamin, N. 2006. Exotic taxa less related to native species are more invasive. Proc. Natl. Acad. Sci. 103, 5841-5845.

Stock, M.W., Castrovillo, P.J. 1981. Genetic relationships among representative populations of five *Choristoneura* species: *C. occidentalis*, *C. retiniana*, *C. biennis*, *C. lambertiana*, and *C. fumiferana* (Lepidoptera: Tortricidae). Can. Entomol. 113, 857-865.

Tan, D.S.H., Ang, Y., Lim, G.S., Ismail, M.R.B., Meier, R. 2010. From 'cryptic species' to integrative taxonomy: an iterative process involving DNA sequences, morphology, and behaviour leads to the resurrection of *Sepsis pyrrhosoma* (Sepsidae: Diptera). Zool. Scripta. 39, 51–61.

Tay, W.T., Behere, G.T., Batterham, P., Heckel, D.G. 2010. Generation of microsatellite repeat families by RTE retrotransposons in lepidopteran genomes. BMC Evol. Biol. 10, 144.

Van Oppen, M.J.H., McDonald, B.J., Willis, B., Miller, D.J. 2001. The evolutionary history of the coral genus *Acropora* (Scleractinia, Cnidaria) based on a mitochondrial and a nuclear marker: reticulation, incomplete lineage sorting, or morphological convergence? Mol. Biol. Evol. 18, 1315-1329.

Van Valen, L. 1976. Ecological species, multispecies, and oaks. Taxon. 25, 233–239.

Volney, W.J.A., Fleming, R.A. 2007. Spruce budworm (*Choristoneura* spp.) biotype reactions to forest and climate characteristics. Glob. Change Biol. 13, 1630-1643.

Walsh, A.M., Kortschak, R.D., Gardner, M.G., Bertozzi, T., Adelson, D.L. 2012. Widespread horizontal transfer of retrotransposons. Proc. Natl. Acad. Sci. 110, 1012-1016.

Wang, X.-P., Yang, G.-J. 2008. A new species of *Choristoneura* Lederer, with a key to the species from China (Lepidoptera: Tortricidae: Tortricinae). Zootaxa. 1944, 66-68.

Weber, U.M., Schweingruber, F.H. 1995. A dendroecological reconstruction of western spruce budworm outbreaks (*Choristoneura occidentalis*) in the Front Range, Colorado, from 1720 to 1986. Trees. 9, 204-213.

Wheeler, Q.D. 2008. The New Taxonomy. CRC Press, NY, USA.

Wheeler, Q.D., Meier, R. 2000. Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, USA.

Witt, J.D.S., Hebert, P.D.N. 2000. Cryptic species diversity and evolution in the amphipod genus *Hyalella* within central glaciated North America: a molecular phylogenetic approach. Can. J. Fish. Aquat. Sci. 57, 687-698.

Yoon, C. K. 2010. Naming Nature. W.W. Norton & Company, Inc., New York, NY, USA.

Zeuner, F.E., 1943. Studies in the systematics of *Troides* Hübner (Lepidoptera Papilionidae) and its allies: distribution and phylogeny in relation to the geological history of the Australasian Archipelago. Trans. Zool. Soc. London. 25, 107–184.

Chapter 2

Phylogenomics of the spruce budworm species complex (*Choristoneura fumiferana*)

2.1. Introduction

New molecular genetic technologies such as next generation sequencing (NGS) have allowed investigation of ecological and evolutionary problems on a finer-scale than ever before (Andrew et al., 2013). The massive data output provided by NGS allows efficient access to genetic divergences (Williams et al., 2010), evolutionary relationships of species (Wagner et al., 2013; McCormack et al., 2012; Jones et al., 2013), population structure and landscape genetics (Emerson et al., 2010; Gompert et al., 2010; Zellmer et al., 2012), dispersal and hybridization (Hohenlohe et al., 2011), microbiome diversity (Xie et al., 2010), and the evolution of adaptive traits and their genetic basis (Colbourne et al., 2011). One approach called genotyping-by-sequencing (GBS) uses NGS to quickly and inexpensively sequence thousands of DNA fragments associated with restriction sites across the genomes of multiple specimens (Elshire et al., 2011; Davey and Blaxter 2010; Davey et al., 2011). The high density of restriction sites throughout genomes provides an excellent basis for phylogenomics, the systematic analysis of genome-wide samples of characters (Meusemann et al., 2010). Through analysis of shared restriction sites, the genome-wide coverage of GBS can be used to determine intraspecific population structure and resolve the phylogenies of closely related species (Cariou et al., 2013; Jones et al., 2013).

One such group of closely related species is the spruce budworm (SBW) complex, also known as the *Choristoneura fumiferana* (Clemens 1865) group. There are 38 species in the genus *Choristoneura* world-wide (Brown et al., 2005), of which 17 are Nearctic (Wang and Yang, 2008). Eight Nearctic coniferophagous species are currently recognized as comprising the SBW complex (Brown et al.,

2005), and these are further divided in to 15 biotypes (Volney and Fleming, 2007). The eight SBW species are major forest pests on conifers (Volney and Fleming, 2007) and include *C. biennis* Freeman 1967, *C. carnana* (Barnes and Busck 1920), *C. fumiferana*, *C. occidentalis* Freeman 1967, *C. orae* Freeman 1967, *C. lambertiana* (Busck 1915), *C. pinus* Freeman 1953, and *C. retiniana* (Walsingham 1879). *Choristoneura fumiferana* and *C. pinus* are sympatric across the boreal forest of Canada and the north-eastern U.S.A., while the remaining six are sympatric or parapatric across the western side of the continent (Lumley and Sperling, 2011a). The eight SBW species are often difficult to discriminate, both visually and ecologically, because morphological characters are highly variable within and among species (Powell and De Benedictus, 1995) and some species share host plants (Lumley and Sperling, 2011a). The original descriptions by Freeman (1953; 1967) acknowledged that their distinguishing characters (morphology, colour, distribution, and food plants) were insufficient to reliably identify all specimens (Freeman, 1967). However, differences in sexual isolation (Campbell, 1967), pheromones (Sanders, 1974; Silk and Kuenen, 1988), and life history characteristics (Harvey, 1967) support species level designation for some biotypes. Overall, the complex evolutionary relationships of the group demonstrate the need for a fine-scale investigation of characters throughout the genome, such as those provided by GBS.

Previous phylogenetic analyses of SBW species based on isozymes (Stock and Castrovillo, 1981; Harvey, 1985; Harvey, 1996), microsatellites (Lumley and Sperling, 2011a), mitochondrial cytochrome c oxidase I (COI) and COII genes (Sperling and Hickey, 1994; Sperling and Hickey, 1995; Lumley and Sperling, 2011a), and COI plus part of the nuclear 28S gene (Dombroskie, 2011) have produced discordant topologies (Fig. 2-1). In particular, the internal relationships of the western species and the position of *C. pinus* remain uncertain. However, NGS-assisted molecular character analysis, in addition to morphology and ecology, has the potential to illuminate their evolutionary history, delimit species, and quantify divergence. With genome-wide GBS sampling of thousands of single

nucleotide polymorphisms (SNPs), it should be possible to resolve the molecular phylogeny of this species group more definitively.

The aims of this study are: a) to evaluate the usefulness of GBS in delimiting the species of the SBW complex, and b) to determine the phylogenetic relationships of these species, with particular focus on the origin of *C. pinus*. Six of the eight SBW species (Brown et al., 2005) and two outgroup species within *Choristoneura* were sampled across their respective ranges (Lumley and Sperling, 2011a). Particular emphasis was placed on collections within Alberta, Canada, because of the potential for species interactions due to overlapping ranges in this region. DNA sequences generated by NGS were mined for SNPs using GBS, and analysed using phylogenetic, population genetic, and phylogeographic approaches.

2.2. Methods

2.2.1. Sample collection

DNA from previously collected *Choristoneura* specimens, stored at -70°C as DNA or adult moths, was used for the first round of GBS analysis with the ApeKI restriction enzyme (Table A-1; *C. fumiferana* N=42, *C. pinus* N=8, *C. occidentalis* N=28, *C. biennis* N=8, *C. carnana* N=8, and *C. retiniana* N=3). Specimens were selected to maximize SBW species diversity, geographic range and host associations across Canada and the United States (Figs. 2-2, and 2-3). Two other widespread North American species of *Choristoneura* served as outgroups (*C. rosaceana* (Harris 1841) N=3, and *C. conflictana* (Walker 1863) N=2). These two species feed on deciduous tree foliage and are broadly sympatric with the coniferophagous *C. fumiferana* group (Vakenti et al., 2001; Lindroth, 1991).

Due to limited read depth from the ApeKI analysis caused by the large number of restriction sites, even after obtaining two rounds of Illumina sequencing, a second GBS analysis was carried out using another restriction

enzyme combination (PstI-MspI). This analysis also focussed on including additional samples from Alberta and Saskatchewan (Fig. 2-3; *C. biennis* N=6, *C. carnana* N=1, *C. conflictana* N=2, *C. fumiferana* N=59, *C. occidentalis* N=8, *C. pinus* N=65, and *C. rosaceana* N=3). Newly collected specimens were identified to species on the basis of their collection location, host plant associations (Lumley and Sperling, 2011a; Freeman 1967), and phylogenetic association. Phylogenetic association was determined from eighteen specimens replicated in the first and second GBS analyses (*C. biennis* N=2, *C. carnana* N=1, *C. conflictana* N=1, *C. fumiferana* N=4, *C. occidentalis* N=4, *C. pinus* N=4, and *C. rosaceana* N=2). Sampling methods employed larval collections from host trees, pheromone traps (Contech lures for *C. fumiferana* cat# 300000092 or *C. pinus* cat# 300000194), and ultraviolet (UV) or mercury vapour (MV) light traps and sheets (Table A-1). Foliage of each host plant was collected along with the larvae and used to rear them to the adult stage. Moths were frozen and stored at -70°C after their meconium was expelled.

2.2.2. DNA extractions and purification

DNA extraction methods for previously collected specimens are detailed in Sperling and Hickey (1994) and Lumley and Sperling (2010). DNA from new specimens was extracted from the whole thorax, legs, and, for smaller specimens, the first one or two segments of the abdomen using a DNeasy blood and tissue spin column kit (Qiagen, Victoria, British Columbia, Canada). The head, wings, and remainder of the abdomen were retained as vouchers at the University of Alberta Strickland Museum and stored at -70°C.

All DNA was purified using ethanol precipitation. DNA purity and quality was assessed using spectrophotometry (NanoDrop 1000, Wilmington, Delaware, USA), and fluorescence (Qubit® 2.0 dsDNA BR assay kit, Invitrogen, Burlington, Ontario, Canada). Only samples with 260/280 nm absorbance ratios between 1.61 and 2.11, and 260/230 ratios between 0.97 and 2.73, were sequenced.

2.2.3. Genome reduction and Genotyping by Sequencing

For each specimen analyzed using the ApeKI restriction enzyme (Fig. 2-4), 300ng of DNA was shipped in 10 ng/µl solutions to Institut de Biologie Intégrative et des Systèmes (IBIS) at Université Laval, Quebec. The IBIS facility used 100ng per library preparation according to the protocol outlined in Elshire et al. (2011). The procedure in brief was as follows: a) using a 96 well plate, each well received a unique barcode adapter (Elshire et al., 2011), common adapter, and single individual's DNA; b) individual DNA samples were digested with the ApeKI restriction enzyme (5' G*CWGC 3') for 2 hours at 75°C; c) DNA fragments were ligated to adapters using T4 ligase; d) samples were multiplexed in groups of 48 and purified to remove adapter dimers; e) adapter-DNA combinations were then PCR-amplified (Elshire et al., 2011), also incorporating an oligonucleotide sequence for binding to the Illumina flowcell; f) PCR product was purified to remove unincorporated primers and nucleotides, and size quantified to select DNA fragments 170-350 bp in length for sequencing. It should be noted that DNA fragments that ligated to two of the same adapters formed a hairpin and were not amplified. In addition, ApeKI is methylation sensitive, meaning it will not cut recognition sequences that have 5-methylcytosine on the 3' base of both strands. This inherent filter ensures that repetitive fractions of the genome (Elshire et al., 2011) were not amplified. The library preparation was sequenced first in a test run of the GBS procedure using single-end sequencing on Illumina HiSeq2000 (Illumina Inc., San Diego, California, USA). However, the read depth of individual loci was low, so the lanes were pooled and sequenced a second time.

For each specimen analyzed using the PstI-MspI restriction enzyme combination, 200ng DNA was shipped in 20 ng/µl solution to IBIS. Libraries were prepared according the protocol outlined in Poland et al. (2012). This restriction enzyme combination was chosen to: a) maximize data compatibility with parallel analyses in other projects; b) sequence fewer fragments to provide greater read depth for each locus; and c) compare results from different enzymes.

In addition to the procedure outlined above, a one nucleotide complexity reduction was used, effectively reducing the number of possible loci to a quarter in order to gain greater read depth (Sonah et al., 2013). A duplex specific nuclease treatment was also applied because some of the restriction sites may have been located in repeated elements that may monopolize many reads and reduce the read coverage for other loci. Duplex specific nuclease selectively removes sequences originating from repeated elements, and is used to normalize cDNA libraries (Zhulidov et al., 2004).

We predicted the number of restriction sites by searching for the recognition sequence in a *C. fumiferana* reference genome (R. Levesque and M. Cusson, unpublished), using standard calculations (number of sites = site frequency x genome size) given a genome size of approximately 530 Mb (M. Cusson, pers. comm. July 26, 2012; genome size determinations carried out by S. Johnston). The site frequency was calculated using the recognition sequence of PstI (CTGCAG) or ApeKI (GCTGC + GCAGC) and GC content of the reference genome (GC ratio = 0.38; B. Brunet, pers. comm. August 16, 2012). In contrast, the MspI restriction enzyme (CCGG) cuts more frequently, so the number of restriction sites for the PstI-MspI digest was calculated using the less frequent cutter. Also, the one nucleotide complexity reduction lowered the number of sites to a quarter of the original number for PstI-MspI.

2.2.4. Filtering sequences and the TASSEL pipeline

The sets of sequences associated with ApeKI and PstI-MspI restriction sites were analyzed separately using the TASSEL (Trait Analysis by aSSociation Evolution and Linkage) pipeline (Bradbury et al., 2007) on the computational resources provided by Compute Canada WestGrid (Fig. 2-4). In order to retain only high quality sequences, the raw reads were trimmed to 64 bases, removing the end where sequencing errors are more likely to occur (Glenn, 2011). Furthermore, reads containing an "N" in the 64 bases were discarded. The barcode portion was also removed but the restriction site overhang was retained.

Reads were further truncated and the length recorded if one of the following was detected: a) the common adapter sequence due to a DNA fragment length of less than 64 bases, or b) the restriction enzyme recognition sequence due to a partial digestion or ligation of two DNA inserts.

Reads from all specimens were pooled together, retaining sequences with two or more reads, which created a master list of sequences. In this way, singletons were discarded to remove potential read errors, but alleles with low read depth were kept, which was important considering the low read coverage for ApeKI even with the two sequencing runs pooled (Table 2-1). It was important to maximize reads at this stage because SNPs from overlapping reads were merged in a later step, and deleting reads early may remove alleles, artificially decreasing the level of heterozygosity. Other researchers (McCormack et al., 2012; Sonah et al., 2013) have used similar settings of 3 and 2 reads/allele respectively. We ran parallel analyses using cut-off levels of 5, 10, and 15 reads/sequence to test for inclusion of null alleles and effect on proportion of heterozygous base calls.

This master list of sequences was aligned against the reference genome of *Choristoneura fumiferana* (560,420 contigs; April 19, 2011), supplied by the labs of Roger Levesque (Institut de biologie intégrative et des systèmes, Université Laval, Quebec) and Michel Cusson (Natural Resources Canada, Laurentian Forestry Centre, Quebec) (unpublished), using Burrows-Wheeler alignment (BWA) tool (Li and Durbin, 2009) on default settings. We had the option of using a reference genome composed of contigs or of scaffolds, which consisted of a subset of the contigs pieced together. We used the reference genome of contigs because it consistently produced more SNPs than the reference genome of scaffolds that was also available (30517 scaffolds, April 19, 2011) under a variety of parameters. Only sequences that aligned to a unique position proceeded to further analysis.

After the alignment positions of all sequences were obtained, the next step in the TASSEL pipeline was to count the occurrence of each sequence per specimen (Bradbury et al., 2007). This was done by revisiting the raw reads files using the same protocol as above but padding sequences shorter than 64 bases

with poly-As. The sequences of each specimen were compared to the reference genome to find SNPs. These SNPs were filtered to remove gaps or tri- and tetra-allelic loci, because these genotypes are considered likely to have been caused by a sequencing or alignment error (Glaubitz et al., 2012). They were further filtered to remove SNP loci with a minor allele frequency less than 0.02, as recommended for unrelated individuals (Glaubitz et al., 2012). The TASSEL pipeline design required the minimum minor allele count threshold to be close to the total number of specimens (90 for ApeKI, and 130 for PstI-MspI) in order to utilize the minimum minor allele frequency filter.

Duplicate SNP loci, from sequences overlapping the same position on the genome, were merged if they shared the same pair of alleles and the mismatch rate for that SNP over all specimens was not greater than two (number of specimens for which the two genotypes disagree/ number of specimens for which neither SNP has missing data) as recommended by the TASSEL project manager for data with low read depth and high heterozygosity (J. Glaubitz, pers. comm. January 23, 2012). Duplicate SNP loci above the mismatch rate threshold were deleted because they were likely to be paralogous loci. If a SNP was merged, then individuals with genotypic disagreements were called heterozygotes. After this, SNPs with a minor allele that occurred in less than 0.02 of the specimens genotyped were, again, removed. Matrices were then created for all SNPs genotyped in greater than 0% (all SNPs), 10%, 20%, 30%, 40%, 50%, 60%, 70%, 75%, 80%, 90%, 95%, and 99% of the specimens. Specimens that were genotyped at less than 10% of the total loci were excluded manually.

2.2.5. Filtering sequences and the UNEAK pipeline

The UNEAK (Universal Network-Enabled Analysis Kit) pipeline is similar to the TASSEL pipeline except it does not align sequences to a reference genome (Lu et al., 2013). It aligns each sequence to other sequences pairwise, and a one bp mismatch is considered a candidate SNP. During alignment a 0.03 error tolerance rate (default) was used. These tag pairs form networks, which are

filtered to remove complicated networks, likely composed of repeats, paralogs, or errors (Lu et al., 2013). The remaining simple networks of reciprocal tag pairs are used to call SNPs in each specimen. The resulting SNPs were filtered using a minimum minor allele frequency of 0.05 and maximum minor allele frequency of 0.5 (defaults). SNP occurrence matrices were created for SNPs genotyped in greater than 10% and 75% of the specimens.

2.2.6. SNP matrix manipulation, diagnostic SNPs, and mtDNA SNPs

Each SNP matrix was concatenated into a continuous sequence for each specimen. The number of base calls were counted (A, C, G, T, R, Y, K, M, S, W, and N=missing data) and the proportions of heterozygous calls were calculated using a custom Perl script. From the matrices containing only SNPs with 75% or greater coverage across all specimens, SNPs were counted as diagnostic if they had a genotype that consistently occurred in one species and no other species (*e.g.* if all *C. pinus* had an A and no specimens from other species had an A, M, R, or W at that locus, N was ignored). Since female specimens of *C. fumiferana* were observed to cluster together in later analyses, this all-female *C. fumiferana* clade was also assessed for diagnostic SNPs. The single contig containing the diagnostic female SNP was BLASTn searched against the NCBI database.

To determine if any SNPs were positioned in the mitochondrial genome, we first determined which contigs of the *C. fumiferana* reference genome were mitochondrial. *Choristoneura fumiferana* contigs were formatted as a searchable database and cross-referenced with the entire mitochondrial genome of the apple leaf roller (*C. longicellana* (Walsingham 1900), GenBank accession HQ452340.1 (Zhao and Zhu, 2012)). The apple leaf roller is the closest relative of the SBW with a mitochondrial genome available (Regier et al., 2012). Cross-referencing was performed using a local_blast_client.pl script provided by P. Stothard (University of Alberta, Dept. AFNS, pers. comm. June 6, 2011). The mitochondrial genomes of three other Tortricidae were also searched against the *C. fumiferana* reference genome (*Adoxophyes honmai* Yasuda 1998

(DQ073916.1) (Lee et al., 2006), *Grapholita molesta* (Busck 1916) (HQ116416.1) (Gong et al., 2012), and *Spilonota lechriaspis* Meyrick 1932 (NC_014294.1) (Zhao et al., 2011)). Matching contig sequences were mapped together into a pseudo-mitochondrial genome, and the specific positions of SNPs within mitochondrial genes were estimated using the *C. longicellana* sequence.

2.2.7. Phylogenetic analysis of SNP matrices

All trees were rooted with *C. rosaceana* and *C. conflictana*. Maximum parsimony (MP) and neighbour joining (NJ) analyses were performed on the ApeKI and PstI-MspI SNP sets that contained SNPs with at least 10% and 75% coverage, as derived from the TASSEL pipeline and also from the UNEAK pipeline. From the TASSEL results NJ trees were also built from the 50% and 90% SNP sets to test the effect of missing data on phylogenetic resolution. Some SNPs were phylogenetically uninformative, where the minor allele was genotyped in only one specimen. If these SNPs were genotyped in very few specimens, then they passed the minimum minor allele frequency filter. However, these phylogenetically uninformative SNPs were excluded from the MP search. Loci with missing data were retained, and a MP topology was inferred from 1000 bootstrap replicates (Felsenstein, 1985) of a heuristic search using subtree-pruning-regrafting (SPR) (Nei and Kumar, 2000) from 10 initial trees formed using random addition. The NJ phylogram was inferred from 1000 bootstrap replicates using a p-distance model that included all substitutions (both transitions and transversions), assuming uniform rates among sites, a homogeneous pattern among lineages, and partial deletion of sites with greater than 95% missing data. Both MP and NJ were performed using MEGA5 (Tamura et al., 2011). Maximum likelihood (ML) was performed on the ApeKI and PstI-MspI SNP sets containing only SNPs with 75% coverage or greater, using RAxML (Stamatakis, 2006a; Silvestro and Michalak, 2012) with 100 rapid bootstrap replications and a thorough ML search under the general time reversal CAT approximation model (GTRCAT). This model is appropriate for large datasets with more than 50 taxa

(Stamatakis, 2006b; Stamatakis, 2008). Branch lengths were recorded for later comparisons of divergences.

Bayesian inference was performed using MrBayes (Ronquist et al., 2012) with a subset of the specimens (Table 2-2) from the ApeKI and PstI-MspI 75% coverage datasets, because the computational cost of running all specimens was beyond the capabilities of available computer clusters. Specimens were chosen to maximize representation of the geographical distributions of species and the diversity within different phylogenetic clades as determined using other phylogenetic programs. The GTR invgamma evolutionary model was run for five million generations for both datasets, at which point convergence was reached as indicated by average standard deviations of split frequencies below 0.01 and potential scale reduction factor (PSRF) values close to 1 (Ronquist et al., 2011). Results were collected with a relative burn-in of 0.25. This subset of specimens was also examined using ML, MP, and NJ in order to compare the Bayesian inference topology to an expected topology. Maximum likelihood was analysed on RAxML with 100 rapid bootstrap replicates under GTR GAMMA, and MP and NJ on MEGA5 with 1000 bootstrap replicates under the p-distance model with 95% partial deletion for NJ and the subtree-pruning-regrafting search algorithm for MP.

The COI gene had previously been sequenced for 27 ApeKI specimens (7 *C. fumiferana*, 10 *C. occidentalis*, 2 *C. biennis*, 4 *C. carnana*, 3 *C. retiniana*, and 1 *C. conflictana*; Lumley and Sperling, 2011a). To compare the SNP and COI phylogenies, we produced MP, NJ, and ML phylogenies of these specimens using SNP data and also using COI data. Maximum parsimony and NJ phylogenies were produced using MEGA5 with 100 bootstrap replicates, under the p-distance model for NJ, and the subtree-pruning-regrafting search algorithm for MP. Maximum likelihood was analysed on RAxML with 100 rapid bootstrap replicates under GTR GAMMA.

2.2.8. Isolation by Distance, inter and intraspecific genetic distances

The pairwise p-distances (#differences / #sites compared) of the 10% and the 75% ApeKI and PstI-MspI SNP sequences were calculated for each specimen using MEGA5 (Tamura et al., 2011), including all substitutions, assuming uniform rates, and assuming a homogenous rate pattern among lineages. The 10% threshold was chosen to include the most information possible, and the 75% threshold was chosen as a comparison in the event that null alleles in the 10% produced too much noise, relative to signal. Pairwise distances were corrected to include the average number of invariant bases per SNP (53 for ApeKI, 73 for PstI-MspI), calculated from the proportion of SNP loci to total nucleotides in the sequences associated with each SNP. By correcting the p-distance we were able to compare our results to other DNA sequence-based studies. The pairwise geographical distance (km) was calculated from the latitude and longitude of collection locations using the Geographic Distance Matrix Generator (Ersts, 2013). A linear regression of pairwise genetic and geographic distances was produced for each species or species group. Average intraspecific and interspecific p-distances were calculated for each dataset. Significance was calculated using a two-sample t-test that compared intraspecific distances to interspecific distances of each species and species comparison.

2.3. Results

2.3.1. Descriptive Genotyping by Sequencing results

Sequencing runs for the ApeKI library produced 457 million reads across 102 specimens, and those for the PstI-MspI library produced 554 million reads for 144 specimens (Table 2-1). The *C. fumiferana* genome had 632,335 ApeKI and 85,624 PstI restriction sites, based on our direct search of the reference genome (contigs, April 19, 2011, unpublished), while we estimated 428,235 ApeKI and 66,376 PstI sites using calculations based on genome size and nucleotide content

(Table 2-1). However the actual number of PstI-MspI sites should be 16,594, one quarter of the predicted value, because of the one nucleotide complexity reduction (Table 2-1).

2.3.2. SNP filtering using the TASSEL and UNEAK pipelines

Fewer SNPs were found using the UNEAK pipeline, when the sequences were not aligned to a reference genome (Table 2-3), compared to the number when sequences were aligned to a reference sequence using the TASSEL pipeline (Fig. 2-5). The TASSEL pipeline produced 789,600 SNPs from ApeKI, and 201,748 from PstI-MspI at a read depth threshold of 2 and a locus filter of 10% (Fig. 2-5). In contrast, the UNEAK pipeline produced 42,262 SNPs from ApeKI, and 16,543 SNPs from PstI-MspI, at the same filter levels (Table 2-3). The locus filter removes loci with missing data, according to the minimum proportion of specimens genotyped. To focus on the dataset with the most information, the remainder of this section describes the TASSEL pipeline results. In either pipeline, if a read is sequenced without a barcode it cannot be identified with its specimen. Almost all of the reads from the PstI-MspI GBS run contained a barcode and restriction site (0.967), but slightly less than half (0.477) of the ApeKI reads did (Table 2-1). ApeKI likely produced poorer results because the initial run was a test of a new library preparation and sequencing platform. After collapsing the reads into unique sequences, there was an average of 5.7 reads per unique sequence for the ApeKI data and 10 reads per unique sequence for PstI-MspI. However, many of the sequences had only one read, and this proportion was slightly higher for ApeKI (0.824) than for PstI-MspI (0.812). With these singletons removed, the average read depth improved to 27.8 reads per sequence in ApeKI and 49 reads per sequence in PstI-MspI.

About half of the unique sequences (54% ApeKI, 47% PstI-MspI) with two or more reads aligned to a unique position on the *C. fumiferana* reference genome (Table 2-4). Roughly a tenth (12% ApeKI and 8% PstI-MspI) aligned to multiple positions on the genome, and the remaining 34% and 46% were

unaligned. The proportion of sequenced reads that aligned to unique positions on the reference contigs was consistent between restriction enzymes. It was also consistent between minimum read depths of 2, 5, 10, or 15.

The number of SNPs and the number of contigs decreased as the minimum proportion of specimens genotyped (locus filter) became more stringent (Fig. 2-5). However the proportion of heterozygous base calls continued to increase until the proportion of specimens genotyped exceeded 95%. This pattern held true for both ApeKI and PstI-MspI sets across all minimum read depths; however the increase in heterozygous base calls was greater for the ApeKI-associated sequences.

Three ApeKI specimens had average locus genotype coverage of less than 10% and were removed from further analysis (*C. fumiferana* A45, and *C. occidentalis* A49, A62). *Choristoneura fumiferana* specimens, the same species as the reference genome, averaged the highest genotyping coverage of all species included in the ApeKI or PstI-MspI analyses (Fig. 2-6). Other species in the spruce budworm species group had slightly lower genotyping coverage, followed by the outgroup species, *C. rosaceana* and *C. conflictana*. The overall genotype coverage was higher in the PstI-MspI analysis (0.514) than for ApeKI (0.415) specifically at a read depth threshold of 2 and 10% locus genotype coverage per specimen (Fig. 2-6), but this held true for other read depth thresholds and loci filter levels.

2.3.3. Diagnostic SNPs

The number of diagnostic SNPs found for a group (species or species combinations) was related to the number of specimens in that group, as well as the amount of genetic divergence between the groups. Groups with fewer specimens often produced more diagnostic SNPs, presumably because unique genotypes are more likely to be found by chance for a few individuals than for a large number of individuals (Table A-2). Despite this bias, *C. pinus* had more diagnostic SNP genotypes than any other coniferophagous *Choristoneura* species, for both the

ApeKI (N=8) and PstI-MspI (N=65) datasets. The western species, *C. occidentalis*, *C. biennis*, and *C. carnana*, had few diagnostic SNPs (Table A-2).

One SNP consistently distinguished female *C. fumiferana* specimens (N=22) from males (N=34) in the PstI-MspI dataset. The contig that contained this SNP (contig # 219743, 660 bp in length, locus position 491, G-C alleles) had a poor-quality match in the NCBI database to a *C. fumiferana* parasitoid virus sequence. The same section of sequence also matched an odorant receptor of *C. rosaceana*.

2.3.4. Mitochondrial DNA search results

Four ApeKI SNPs were found in mitochondrial DNA (mtDNA), out of the 789,628 SNPs total in the read depth threshold of 2 and 10% locus genotype coverage set. They were all within 43 base pairs of each other and near the end of the NADH dehydrogenase subunit 5 coding sequence (Fig. 2-7). The first two were synonymous, occurring at the third codon position of the 312th and 313th amino acids (proline and alanine, respectively). The other two SNPs were non-synonymous, occurring in the second base position of the 325th amino acid (lysine/arginine) and the third base position of the 326th amino acid (methionine/isoleucine).

Three PstI-MspI SNPs were found in mtDNA, of the 201,791 SNPs in the read depth threshold of 2 and 10% locus genotype coverage set. They were within eight bases of each other and appeared to be synonymous. One occurred on the third base pair of the second last amino acid (asparagine) of the ATP6 coding sequence, and the other two occurred between the ATP6 coding sequence and the COX3 coding sequence. Neither the ApeKI nor PstI-MspI mitochondrial SNPs distinguished any species.

2.3.5. Phylogenies based on SNP sets for each restriction enzyme

The NJ phylogenies constructed from SNPs with 90% or more genotype coverage produced less resolution than the phylogenies using a 50% threshold (data not shown). This pattern was seen for both the PstI-MspI and ApeKI analyses. The most resolved phylogenies from both restriction enzyme sets were derived using the largest dataset of loci with 10% or more genotype coverage.

Phylogenetic analyses of the ApeKI and PstI-MspI SNP datasets placed *C. fumiferana* as the sister lineage to the clade of western species (*C. occidentalis*, *C. biennis*, *C. carnana*, and *C. retiniana*), and *C. pinus* basal to this (Figs. 2-8, and 2-9). In the PstI-MspI analysis, female *C. fumiferana* (N=22) formed a clade separate from male *C. fumiferana* (N=34), however, this was not the case for the ApeKI analysis (only 4 females and 3 males sexed; Fig. A-1). A potential hybrid specimen collected in southern Alberta was topologically located basal to the *C. fumiferana* clade, close to the *C. occidentalis* clade in the PstI-MspI analysis. This specimen was collected on Colorado blue spruce.

Within the western clade, *C. retiniana* constituted a strongly supported monophyletic clade that was basal to the other species in the group (Fig. 2-8). *Choristoneura occidentalis* was split in two distinct clades; with one strongly supported monophyletic clade consisting of specimens from Arizona and Nevada (Figs. 2-8, and 2-9). The remainder of the *C. occidentalis* specimens were from Alberta, British Columbia, and Montana, and were polyphyletic with respect to *C. biennis* and *C. carnana*. *Choristoneura biennis* was polyphyletic in the ApeKI analysis, with one clade of specimens from British Columbia, and the Alberta specimens scattered among the northern *C. occidentalis* specimens. This topology was also found in the COI results of the 27 ApeKI specimens, with the exception that *C. retiniana* was not basal to the other western species, rather it was monophyletic within the *C. occidentalis* clade. In the PstI-MspI analysis, the *C. biennis* specimens were monophyletic. The *C. carnana* specimens, all from California, formed a monophyletic clade within the northern *C. occidentalis* clade.

There was very little bootstrap support for population structuring within species other than the clades mentioned above (Figs. 2-8, and 2-9).

The phylogenetic analyses of SNPs from the UNEAK pipeline produced the same topology as SNPs from the TASSEL analysis in almost all cases. These phylogenies used SNPs that passed the 75% loci genotype filter, and did not use the reference genome to mine SNPs. The MP and NJ analysis of the ApeKI SNPs, and ML and NJ of the PstI-MspI SNPs, placed *C. fumiferana* and the western species clade as sister taxa, with 61%, 92%, 96%, and 79% bootstrap values respectively. Individual monophyly of *C. pinus*, *C. fumiferana*, *C. retiniana*, and the *C. occidentalis-biennis-carnana* clade were all strongly supported with bootstrap values over 80%.

Occasionally, alternate topologies were produced from the UNEAK pipeline SNPs. Under ML the ApeKI SNPs placed *C. pinus* and *C. fumiferana* as sister taxa with 40% bootstrap support. Similarly, under MP the PstI-MspI SNPs placed these species together with no bootstrap support, but only at the 75% locus filter. At the 10% locus filter, the more common topology with the *C. fumiferana* and western species clade was retained with 91% bootstrap support. Similarly, Bayesian inference of the TASSEL ApeKI and the TASSEL PstI-MspI SNP sets produced an alternate topology. It resulted in a polytomy of the *C. pinus*, *C. fumiferana*, and western lineages, despite reaching convergence, an average standard deviation of split frequencies of 0.00506 for PstI-MspI and 0.003553 for ApeKI, and potential scale reduction factor (PSRF) values at 1 or very close (0.999 to 1.005).

2.3.6. Isolation by distance, inter and intraspecific genetic distances

Intraspecific average pairwise genetic distances (Fig. 2-10, and Table A-3) showed that, within the *C. fumiferana* complex, *C. pinus* had less genetic variation within it than *C. biennis, C. carnana, C. fumiferana,* or *C. occidentalis*. This was true for the ApeKI associated sequences, where the sample size was 8 and all specimens were collected in Ontario, and remained true in PstI-MspI,

where the sample size was 65 with specimens originating from Ontario to Alberta. Overall, the species with the most intraspecific genetic variation was *C. fumiferana*, whereas *C. retiniana*, *C. conflictana*, and *C. rosaceana*, had the lowest levels of genetic variation. However, these species with the lowest genetic variation had the smallest sample sizes and narrow collection ranges (Fig. 2-10).

Despite low intraspecific genetic variation, *C. pinus* had high interspecific genetic distances from the other species of the *C. fumiferana* complex, meaning it was the most genetically divergent SBW (Fig. 2-10, and Table A-4). There was more genetic distance between *C. pinus* and *C. fumiferana* or *C. pinus* and western clade than there was between *C. fumiferana* and the western clade, in both restriction enzyme SNP sets (Fig. 2-10). The western species, *C. biennis*, *C. carnana*, *C. occidentalis*, and *C. retiniana*, had the least genetic divergence from each other. For the PstI-MspI SNPs the difference between the western species intraspecific and interspecific distances was not significant (Table A-4). All other interspecific comparisons were significantly different from their respective intraspecific distances for the PstI-MspI SNPs. Also, in the ApeKI analysis, *C. retiniana* appeared distinct from the other western species (Fig. 2-10). As expected, the outgroup taxa, *C. conflictana* and *C. rosaceana*, were the most genetically different from the ingroups and also from each other (Fig. 2-10).

One specimen in the PstI-MspI phylogenetic analysis was basal to the *C. fumiferana* clade, and topologically between *C. fumiferana* and *C. occidentalis*. When compared to *C. fumiferana* specimens, it returned slightly higher pairwise genetic distances than *C. fumiferana* X *C. fumiferana* comparisons (Fig. 2-10). When compared to *C. occidentalis*, it also returned higher distances than *C. occidentalis* X *C. occidentalis*, but slightly lower values than *C. fumiferana* X *C. occidentalis* pairwise comparisons. Because this specimen seemed to be almost *C. fumiferana* and almost *C. occidentalis*, we provisionally treat it as a hybrid.

There was subtle isolation by distance in *C. occidentalis*, *C. pinus*, and *C. carnana* in both the ApeKI and PstI-MspI analyses. However, *C. biennis* and *C. fumiferana* show isolation by distance in PstI-MspI analysis but not ApeKI

analysis (Fig. 2-11). The results using locus filters of 10% (included loci missing genotypes in up to 90% of specimens) and 75% were almost identical.

2.4. Discussion

2.4.1. Species trees, gene trees, and genome sampling

The best estimate of a species tree is the modal gene phylogeny, because it reflects the evolutionary history of the largest proportion of the genome studied (Edwards, 2009). One advantage of large SNP datasets is that the increased genome coverage has the capacity to produce a better estimate of a species tree than traditional Sanger sequencing. In contrast, phylogenies of single genes are often discordant with each other because they are affected by reversal (Gatesy et al., 2007; Pashley and Ke, 1992), introgression (Bossu and Near, 2009), and horizontal transfer (Walsh et al., 2012). These phylogenies may show the evolutionary history of that single gene but often fail to represent the most frequently occurring modal gene phylogeny. In order to consistently produce an informative modal phylogeny, data from three or more genes is often required (Dupuis et al., 2012; Maddison and Ober, 2011; Corl and Ellegren, 2013). Genes like the commonly used mitochondrial COI and rDNA internal transcribed spacer region 2 (ITS2) had only 36 and 21 phylogenetically informative character sites in the SBW complex (unpublished data from preliminary sequencing, H. Bird). These two were not enough loci to reliably delimit units at the species level, and to delimit units at the finer scales of population and regional level even more loci would be required. For example, in the Beacham et al. (2008) study on Chinook salmon (*Oncorhynchus tshawytscha* Walbaum 1792), the equivalent of 135 informative SNPs were required to delimit units at the population and regional level. However, having genotyped many thousands of loci in this study, we are confident that we have obtained an accurate modal phylogeny for the included species of the *C. fumiferana* complex.

Another advantage of large SNP datasets is that the evolutionary histories of different loci average out. This is because the SNPs were sampled from many coding (and non-coding) regions of the genome, each of which would have undergone slightly different evolutionary histories as a result of variation in selective pressures, chromosomal crossover and linkage (Wolfe et al., 1989). If we had not sampled enough SNP loci, the two enzyme systems would have produced different topologies. But both ApeKI and PstI-MspI produced the same topological groupings. In addition, both topologies had well supported deep branches, and even similar branch lengths (Figs. 2-8, and 2-9). For future research, either enzyme system should provide enough restriction sites and coverage of the genome to produce accurate species phylogenies.

The evolutionary history of some SNP loci are likely concordant with a pattern of introgression, which could add signal that supports alternate topologies. However, the signal of introgression is unlikely to overpower the signal from SNPs following the pattern of successive speciation events. This is because introgression is possible between many SBW species (Sanders, 1974), adding noise instead of causing a directional bias. If there was a directional bias it would likely draw sympatric specimens closer on the tree, as seen in cichlid fish (Rüber et al., 2001), since there is more opportunity for regular contact (Funk and Omland, 2003). In contrast, our SNP phylogenies did not place the two sympatric species *C. pinus* and *C. fumiferana* together as sister species (Figs. 2-8, and 2-9).

The SNP loci producing the signal of successive speciation events also produced a pattern of lineage sorting, which is defined as the convergence of gene phylogenies to the species phylogeny after speciation (Funk and Omland, 2003). Lineage sorting is a result of both genetic drift and divergent selection pressure (Funk and Omland, 2003). Our SNPs were likely randomly distributed through the genome, because they occur in roughly 1 in 7 contigs for ApeKI and 1 in 15 for PstI-MspI (Fig. 2-5). It follows that if 74% to 93% of an invertebrate genome is non-coding (Taft and Mattick, 2003), then the majority of SNPs are likely in non-coding regions, and thus less likely to be under selection. Our SNP sampling is slightly biased towards more conserved regions because only loci associated

with conserved enzyme recognition sites were sampled. However, the majority are still likely to be in non-coding regions and not under selection. This is in agreement with another study which found that the vast majority of SNPs were not under selection pressure in threespine stickleback fish, using NGS (Hohenlohe et al., 2010). Our most resolved phylogeny was produced with the largest number of SNPs, and if the majority of our SNPs are not under selection pressure, then it is likely that the SNPs producing the signal for the species phylogeny are also not under selection pressure. This means that even though divergent selection pressure produces and maintains sympatric speciation (Fitzpatrick et al., 2008) in species like SBWs, it was drift that produced the signal of speciation in the SBW phylogeny. A phylogeny produced by loci not under direct selection should still be accurate, despite introgression and incomplete lineage sorting, if many loci and many individuals per species were sampled (Maddison and Knowles, 2006).

Even though the genetic signal of speciation is produced by genetic drift, species differentiation is maintained by divergent selection pressures (Fitzpatrick et al., 2008). In the SBW species complex where some species are sympatric in parts of their range, closely related, and introgress, selection pressure would play an important role in maintaining species boundaries and genomic integrity (Mallet, 2005), while also promoting introgression of certain advantageous genes (Staubach et al., 2012). Tracing which genes are under divergent selection and the biological innovations they bring to the SBW speciation process is an intriguing project which we begin to pursue in Chapter 3 of this thesis.

2.4.2. The effect of missing data

Restriction enzyme choice is important. The frequency of recognition sites and plexity (specimens per sequencing lane) control the average number of reads per locus. Our first restriction enzyme, ApeKI, was a 4.5mer cutter, with ~500,000 estimated recognition sites, providing much greater coverage of the genome than PstI with ~70,000 sites (or ~17,500 due to the one nucleotide complexity reduction). Both ApeKI and PstI are methylation sensitive (Castel et

al., 2011; Gruenbaum et al., 1981), and because DNA and library preparation procedures do not strip DNA of methylation (Sadri and Hornsby, 1996), recognition sites methylated on the 3' end of both strands were not digested. Methylation is a method of controlling gene expression (Jaenisch and Bird, 2003), so recognition sites that overlap with transcription regulation sites of genes that are turned "off" in an individual were not sequenced, increasing null alleles and variation in genotype coverage.

Distantly related specimens share fewer recognition sites than closely related specimens, which is a weakness of the GBS method critiqued by Arnold et al. (2013). However, the fact that *C. pinus* shared fewer recognition sites with *C. fumiferana* compared to most other species in the group provided further evidence that *C. pinus* was more distantly related to *C. fumiferana* than *C. occidentalis* (Fig. 2-6). The proportion of SNPs genotyped decreased in species more distantly related to *C. fumiferana* (and shared fewer recognition sites) because only sequences that aligned to the *C. fumiferana* genome were analysed. The pattern was more obvious with the ApeKI dataset than the PstI-MspI dataset (Fig. 2-6) partly because small sample sizes decreased the representation of a species' SNPs and therefore decreased genotype coverage. In this case, the western lineage had a low sample size in the PstI-MspI analysis.

Despite shortening the deepest braches of the tree because of polymorphism in the restriction enzyme recognition site, the GBS method produces the correct species topology with strong support, even on deeper branches, as seen in *Drosophila* studies (Cariou et al., 2013). There was a possibility of bias due to increased sequence information in the direction of *C. fumiferana* because only sequences that aligned to the reference genome were included. However, because SNPs from the UNEAK pipeline, which included 40% of sequences that did not align to the reference genome (Table 2-1), produced the same topological groupings as the TASSEL pipeline, this possible bias clearly did not affect the overall topology.

We were concerned that including SNPs with low genotype coverage would incorrectly assign specimens to species clades, similar to how a single gene

may produce an misleading species tree; however we found the opposite to be true. The dataset including SNPs with 10% or more genotype coverage was used for MP and NJ because it resolved more species into monophyletic units than the 50% or 90% datasets. This was because infrequently genotyped SNPs were still phylogenetically informative and greatly improved the level of phylogenetic resolution. Having more loci in a genome-wide analysis of closely related species, despite the increased inclusion of missing data, is known to create a more resolved phylogeny (Wagner et al., 2013; Jones et al., 2013) and, given a large number of SNPs, produces an accurate species phylogeny (Lu et al., 2013).

In order to gain more phylogenetic resolution by including more data, we used the low read depth of two sequences per allele, which may have included more sequencing errors than a higher read depth threshold. However, our analysis pipeline controlled for sequencing errors in multiple ways. Illumina sequencing has one of the lowest error rates, 0.1% (Glenn, 2011) or 0.26% with mostly >Q30 phred scores (Quail et al., 2012), compared to other NGS platforms, and when errors occur they tend to be substitutions (Metzker, 2010) and occur towards the end of the sequence (Glenn, 2011). Our pipeline trimmed the 100bp Illumina sequences to 64bp, removed reads that contained an N, gaps and indels, and also removed singletons (sequences with represented by a single read) which should remove the majority of sequencing errors. Our filters also removed tri- or quad-allelic SNPs and SNPs below a minimum minor allele threshold. So, before we got to the point where we considered removing SNPs with low genotype coverage, we removed most possibilities of read errors. Therefore, the possibility of errors is reduced beyond the point where it should significantly affect our results.

Some SNPs were present in a specimen but were not amplified and genotyped because of a mutation in the enzyme recognition site and are null alleles. Including null alleles decreases the proportion of heterozygous base calls by increasing false homozygous genotypes (Nielsen et al., 2012). Removing low coverage loci increases the proportion of heterozygous base calls, but this increase does not cease as the locus filter reaches its maximum (Fig. 2-5), meaning that

null alleles are present not only among low coverage loci, but also high coverage loci. This should be an expected result, because the GBS technique produces a large sampling variance between alleles (Hohenlohe et al., 2012). We were unable to selectively remove null alleles from our data, but because they occur in all species at different loci, they should not have a directional bias on our results. Additionally, in the PstI-MspI results there is an abrupt decline in heterozygotes when the locus filter reaches 99%. This decline is due to the small number of SNPs at this level, of which few were heterozygotes by chance. In the ApeKI results, there were no SNPs at the 99% locus filter, and so this datapoint was not included.

The proportion of heterozygous base calls in our dataset, 0.04 to 0.12 (Fig. 2-5), was similar to that found in a study comparing closely related teleost fish species (Williams et al., 2010). For example, Williams et al. (2010) reported SNPs with observed heterozygosity ranged from 0.016 to 0.17 among all populations with a mean of 0.1. Similarly, another study using rainbow and westslope cutthroat trout species in danger of hybridizing identified observed heterozygosity peaks between 0 and 0.1 (Hohenlohe et al., 2011). Many of our SNPs were from differences between outgroup taxa and ingroup taxa as can be inferred from the long branches to the outgroups (Figs. 2-8, and 2-9), so it is expected that many loci would be fixed and would be scored as homozygous. However, the lower end of the range of observed heterozygosity of SBWs (Fig. 2-5) is greater than the low end of the fish species (Williams et al., 2010; Hohenlohe et al., 2011).

When comparing enzymes, ApeKI had a greater range in proportion of heterozygous base calls than PstI-MspI (Fig. 2-5). It was lower when including all SNPs because the lower average read depth and genotype coverage caused an over-representation of false homozygotes. However, it was higher than PstI-MspI when including only SNPs of highest coverage because these included many paralogous loci (false heterozygotes) which often have higher coverage. Because ApeKI sampled more loci total, there were more paralogous loci total, and the loci with the highest genotype coverage were likely to be populated by this greater number of paralogous loci.

2.4.3. Challenges of SNP based data and large datasets

Only recently have SNPs been used for distance based phylogenetic analysis (Wagner et al., 2013; McCormack et al., 2012) and maximum likelihood analysis (Eaton and Ree, 2013). Of the distance based models available, the p-distance model calculated the proportion of sites that were different between the two sequences being compared (= #differences / #sites compared), normalizing for missing data (Tamura et al., 2011), whereas the similar "number of differences" model calculates distance as the number of differences divided by sites total (Tamura et al., 2011). Normalizing for missing data was appropriate because we used a relaxed filter of 10% genotype coverage at loci for most phylogenetic analyses, resulting in up to 90% missing data at a locus. We used p-distance for NJ and the DNA substitution model GTR for ML and Bayesian inference, because it was the only model supported by RAxML, the only software able to process our large dataset. The topologies from ML were the same as MP and NJ, so model choice did not have an effect. The polytomy produced by Bayesian inference may have been the result of software design because it is not normally applied to SNP data.

2.4.4. The origin of C. pinus *and mitochondrial genome introgression*

According to our results, *C. fumiferana* and the western clade (*C. biennis*, *C. carnana*, and *C. occidentalis*) in addition to *C. retiniana* were more closely related to each other than to *C. pinus*. This result is in agreement with the isozyme study of Harvey (1996). Harvey (1996) used 12 loci and found the same genetic relationship among these species as we have here. In both his study and ours, *C. pinus* and *C. fumiferana* are monophyletic clades with relatively large genetic distances separating them from the other species. Also, the western species *C. biennis* (Fig. 2-8), and *C. occidentalis* (Figs. 2-8 and 2-9) are polyphyletic with respect to each other and the western lineage is sister clade to *C. fumiferana*. These results agree with studies based on mtDNA (Sperling and Hickey, 1995;

Lumley and Sperling, 2010; Lumley and Sperling, 2011a) in which a well-supported monophyly of *C. pinus* and *C. fumiferana*, and the polyphyly of the western species is evident. However, our results disagreed with the mtDNA phylogeny which places *C. pinus* within the western lineage. In the mtDNA phylogeny, the western lineage has two strongly divergent clades, one sister to *C. pinus*, and a beta haplotype clade between *C. pinus* and *C. fumiferana*. This strong mitochondrial divergence within species brought doubt on the morphology and ecology based species delimitations. However, with the genotyping evidence provided in this thesis, the original taxonomy is supported, and another example of incongruence between mitochondrial and nuclear DNA is discovered.

This discordance between mtDNA and nuclear DNA suggests three possible hypotheses, (1) post-divergence mitochondrial introgression from *C. occidentalis* to *C. pinus*, (2) hybrid origin of *C. pinus*, or (3) *C. pinus* and *C. occidentalis* shared a common ancestor that diverged from *C. fumiferana* and later the *C. pinus* lineage gained enough genetic differences to appear more distantly related to *C. fumiferana* and *C. occidentalis* than they are from each other. The third hypothesis is unlikely because *C. pinus* alone is unlikely to gain more genetic differences than the other two lineages, if most SNPs were located in non-coding regions of the genome and were not under strong divergent selection pressure. The second hypothesis is unlikely because *C. pinus* does not occur as a basal group within one of the "parental clades", *C. fumiferana* or the western lineage, which is how hybrids behave phylogenetically (Triplett et al., 2010). The *C. fumiferana* x *C. occidentalis* hybrid specimen in the PstI-MspI analysis displayed typical hybrid behaviour; basal to the *C. fumiferana* clade, but close to the western lineage. For the first hypothesis to be supported, mitochondrial introgression must have taken place soon after divergence, and/or a selective sweep completely removed the original *C. pinus* mitochondrial lineage, because *C. pinus* is strongly monophyletic in mtDNA studies (Sperling and Hickey, 1995; Lumley and Sperling, 2010; Lumley and Sperling, 2011a). Also in support of the first hypothesis, the genetic distance between *C. fumiferana* and the western lineage is less than the distance between either of them and *C. pinus* (Fig.

2-10), suggesting that *C. pinus* diverged first and *C. fumiferana* and the western lineage shared a more recent common ancestor. Mitochondrial and nuclear phylogenetic discordance signifying mitochondrial introgression has been found in other Lepidoptera (Nice et al., 2002; Gompert et al., 2010) and vertebrates (Bossu and Near, 2009). The mtDNA of *C. retiniana* might have also originated from *C. occidentalis* because COI haplotypes place it within the *C. occidentalis* clade (Lumley and Sperling, 2011a), but SNP results place it basal to the other western species.

Regardless of which species diverged first, the short branch length between the divergence of the *C. pinus* lineage and the following divergence of *C. fumiferana* from the western lineage (Figs. 2-8, and 2-9) suggests that the divergence events of these three main lineages occurred in a short period of time. This was supported by the poor resolution produced when few loci (strict filtering levels) were used. It is also supported by preliminary ITS2 sequencing results (unpublished, H. Bird) which produced a polytomy of the western species (*C. biennis*, *C. carnana*, *C. occidentalis*, *C. orae*, *C. lambertiana*, and *C. retiniana*), and, although it resolved the eastern species (*C. pinus*, and *C. fumiferana*) into their own monophyletic clades, the order of divergence was debatable. Under MP analysis of ITS2 *C. pinus* and the western lineage were sister clades, but under ML analysis *C. pinus* and *C. fumiferana* were sister species.

Although the divergence of the major SBW lineages likely preceded the recession of the Laurentide Ice Sheet, understanding how their host plants recolonized Canada during this recession allows us to speculate on the historic movement of the moths themselves. By dating the commencement of current sympatry we can infer the parallel commencement of possible introgression. White spruce (*Picea glauca* (Moench) Voss), the host of *C. fumiferana,* survived the Wisconsin Glaciation in two refugia on far corners of the continent: Alaska and Kansas-Missouri (Anderson et al., 2011), and was the first to colonize what is now Canada followed by jack pine (*Pinus banksiana* Lambert) in the east, 10k-6k B.P. (Pielou, 1991; Volney, 1985), from its refugia in the Appalachian highlands

and Atlantic coastal plain (Pielou, 1991; Rudolph and Yeatman, 1982). At about the same time as the shift from spruce to jack pine occurred far in the east, Douglas fir (*Pseudotsuga menziesii* (Mirbel) Franco), the host of *C. occidentalis*, was moving north from refugia on both sides of the Rocky Mountains: Washington-Oregon and Wyoming-Utah (Gugger and Sugita, 2010). The convergence of white spruce and Douglas fir in Alberta, Canada, could have brought *C. fumiferana* and *C. occidentalis* in contact with each other, and *C. pinus* would have followed *C. fumiferana* with the same time lag as their host plants, assuming that the moths and trees migrated in parallel.

2.4.5. Alternate topologies, hybrids, and nuclear gene introgression

Support for alternate topologies in our phylogenetic analysis could be the result of convergent selection pressure, incomplete lineage sorting, or introgression. Hybridization is believed to occur in the wild (Lumley and Sperling, 2011a; Lumley and Sperling, 2011b; Harvey, 1996; Nealis, 2005), and one of our specimens appears to be a hybrid between *C. occidentalis* and *C. fumiferana*. It was collected as a larva on Colorado blue spruce in a shelter belt of a homestead in the southern Alberta prairies, together with a non-hybrid *C. occidentalis*. Although both *C. occidentalis* and *C. fumiferana* are known to attack blue spruce (Weber and Schweingruber, 1995; Maine Forest Service, 2000) they have not been documented to do so in Alberta. This supports Volney and Fleming's (2000) idea that recent climate change has allowed populations of SBW to move north and rapidly evolve new host relationships through hybridization. This specimen is likely a backcross to *C. fumiferana* or an advanced generation hybrid because it is more genetically similar to *C. fumiferana* than *C. occidentalis* (Fig. 2-10). A backcross is in agreement with one-way mating attraction found by Sanders et al. (1977) where *C. fumiferana* females attract a higher percentage of *C. occidentalis* males than *C. occidentalis* females attract *C. fumiferana* males. Directional introgression could also be promoted by the increase in egg weights, which is advantageous in colder climates (Harvey, 1983), gained by a female

C. fumiferana outcrossing and by females of other species mating with *C. fumiferana* males (Volney and Fleming, 2007).

2.4.6. Genomic integrities and evolutionary potential

The species *C. pinus*, *C. fumiferana*, *C. retiniana*, and the outgroup species *C. rosaceana* and *C. conflictana* have large numbers of diagnostic or autapomorphic SNPs (Figs. 2-8, 2-9, and Table 2-2). These autapomorphic SNPs are evidence that introgression and hybridization have not completely degraded the genomic integrities of these species. However, there are few unique SNPs for *C. occidentalis*, *C. biennis*, and *C. carnana*. When they are considered together, synapomorphic SNPs are found for them as a group. This raises questions regarding their status as species. Further sampling in parapatric and allopatric areas of their ranges is needed to answer this question.

In further support of the genetic similarity of *C. occidentalis*, *C. biennis*, and *C. carnana*, their intraspecific genetic variation was the same as the interspecific variation between them, whereas *C. fumiferana* and *C. pinus* were more distinct (Fig. 2-10, and Table A-3). The amount of genetic variation within SBW species, 0.08-0.167%, was on the low end of the range of other lepidopteran species. For example, genetic variation ranged from 0.03-2.71% in *Dioryctria*, *Hyles*, more *Choristoneura*, and *Papilio* (Roe and Sperling, 2007), 0.32-0.48% in *Busseola* (Sezonlin et al., 2006), and 0-4.8% in *Dioryctria* (Whitehouse et al., 2011). Compared to a previous study of *Choristoneura* species (Lumley and Sperling, 2010) our results are low, however this could be because our study encompasses whole genome variation instead of only the COI-COII region which is often sequenced because it contains high levels of variation (Roe and Sperling, 2007). A better comparison may be to the whole genome measurements of genetic variation in *Drosophila melanogaster* which ranges from 0-1.0% (Aquadro et al., 2001), or an estimate using multiple genes as in *Bombyx mori*, 0.11% (Karlin and Ladunga, 1994).

The interspecific, but within the genus, variation of 0.08-0.28% is low compared to other lepidopterans, 0.19-2.7% (Brown et al., 1999), 2.56-3.12% (Sezonlin et al., 2006), 0.18-6.76% (Roe and Sperling, 2007), or 0-5.1% (Whitehouse et al., 2011) especially considering our outgroups, which, although within the genus, only diverged 0.29-0.73% from the ingroup. This could be because the more distantly related specimens share fewer restriction enzyme recognition sites which can underestimate the divergence. Comparing our species to each other, we found that *C. fumiferana* had the most genetic variation. The large amount of genetic variation in *C. fumiferana* could be because it was the same species as the reference genome and would have retained the most sequences, and thus the largest amount of variation, compared to the other species. The other species have lower intraspecific genetic variation which could signify that bottlenecks have occurred, possibly due to low points in their cyclical outbreak pattern (Weber and Schweingruber, 1995), or less diversifying selection within the species (Futuyma and Peterson, 1985).

When comparing intraspecific genetic distances to geographic distances we found subtle isolation by distance in *C. pinus*, *C. occidentalis*, and *C. carnana* (Fig. 2-11). There is was little to no isolation by distance pattern for *C. fumiferana* and *C. biennis.* There is also no visible phylogenetically structured organization of genetic variation for *C. fumiferana* in relation to geography. This suggests either a large amount of gene flow across the range of the species, as expected from its large dispersal distances (Sturtevant et al., 2013), or that further filtering of SNPs, or more samples from the extremities of the range are needed to find a clearer signal.

Even though there was no geographical organization within *C. fumiferana*, there was complete separation between the females and the males. There was a single diagnostic SNP that identified the female *C. fumiferana* in the PstI-MspI dataset. The sequence surrounding this SNP matched very poorly to any sequences in GenBank. This is interesting because the females are the heterogametic sex, and would be expected to have a smaller genome size because the W chromosome is typically smaller than the Z chromosome (Yoshido et al.,

2006; Fuková et al., 2005). However, genome sizing results by our colleagues (M. Cusson, pers. comm. July 26, 2012; genome size determinations carried out by S. Johnston; Hare and Johnston, 2011) show that females had significantly larger genomes than males in all species tested; *C. fumiferana*, *C. occidentalis*, *C. pinus*, and *C. rosaceana*. This combined with the poor match from GenBank suggests that the female W chromosome might be full of "junk" DNA from horizontal gene transfer events from parasites and other species, transposons, and gene duplication events, as is the case with the W chromosome in the well annotated *Bombyx mori* genome (Abe et al., 2005; Kawaoka et al., 2011). Another possibility is that the *Choristoneura* W chromosome is a neo-sex chromosome, as many lepidopteran Z chromosomes are known to be (Nguyen et al., 2013), but unlike the Z chromosome, which contains many genes important to speciation (Sperling, 1994), a lack of selective pressure may have allowed it to retain autosomal "junk" DNA.

2.4.7. Conclusions

For species complexes like the North American coniferophagous *Choristoneura*, hybridization and introgression can cause single gene phylogenies to be misleading (Dupuis et al., 2012; Maddison and Ober, 2011; Corl and Ellegren, 2013). Therefore, discovering the true species phylogeny may require many genome-wide markers (Wagner et al., 2013). This makes GBS suitable for phylogenetic analyses of species complexes, where the large number of SNPs provides better resolved trees, despite missing data (Wagner et al., 2013; Jones et al., 2013). Our results placed the jack pine budworm, *C. pinus*, as more distantly related to *C. fumiferana* and *C. occidentalis* than previous mtDNA research predicted (Sperling and Hickey, 1994; Sperling and Hickey, 1995; Lumley and Sperling, 2011a; Dombroskie, 2011). Incongruence between mtDNA and nuclear SNPs suggest *C. pinus* may have had mitochondrial introgression from *C. occidentalis*. Despite this, it is one of the most genetically distinct species in

the *C. fumiferana* group, as demonstrated by its strongly supported monophyly, long branch length, and large number of autapomorphic SNPs.

To expand on this study, future research could include the other two coniferophagous species, the pine-feeder *C. lambertiana*, and the spruce-feeder *C. orae* of the north-western edge of Canada and Alaska. A closer look at the subspecies *C. pinus maritima*, which feeds on *Pinus virginiana* and *P. rigida* in eastern coastal states, and at the species of the western lineage, could determine if these are subspecies, biotypes, or separate species. A broader look at the genus *Choristoneura* could also address questions about the ancestral food source and multiple or single origins of coniferophagy.

Table 2-1
Descriptive Genotyping by Sequencing results from ApeKI and PstI-MspI enzyme digests using the TASSEL pipeline.

	ApeKI	**PstI-MspI**
# specimens	102	144
# reads total	456,867,282	553,617,907
Average # reads per specimen	4,479,091	3,844,569
# barcoded reads total	218,068,779	535,102,441
# unique sequences total	38,124,096	53,453,003
Average # reads per unique sequence	5.72	10.01
# unique sequences with 2 or more reads	6,707,976	10,030,469
Average # reads per sequence with 2 or more reads	27.83	49.02
# unique sequences with 5 or more reads	3,265,071	3,388,794
# unique sequences with 10 or more reads	2,021,933	1,805,581
# unique sequences with 15 or more reads	1,473,593	1,299,131
Expected # of restriction sites[i]	428,235	66,376

[i]Expected number of restriction sites calculated using genome size 530 Mb and proportional GC content 0.38 (genome size determinations carried out by S. Johnston; M. Cusson, pers. comm. July 26, 2012; B. Brunet, pers. comm. August 16, 2012). PstI-MspI sites had a one nucleotide complexity reduction = 16,594. Please note that the number of unique sequences is expected to greatly out number restriction sites, and also most unique sequences do not occur in all specimens, both due to variation in restriction sites and SNPs.

Table 2-2
Bayesian inference subset of specimens from ApeKI and PstI-MspI analyses.

Species	ApeKI	PstI-MspI
C. biennis	2 (A5, A8)	2 (PM66, Pm67)
C. carnana	2 (A11, A16)	1 (PM68)
C. conflictana	1 (A18)	1 (PM69)
C. occidentalis	4 (A79, A81, A82, A83)	4 (PM70, PM71, PM72, PM73)
C. fumiferana	3 (A25, A34, A56)	5 (PM78, PM79, PM81, PM 111, PM112)
C. pinus	2[i] (A86, A87)[ii]	4 (PM59, PM60, PM63, PM64)
C. retiniana	2 (A91, A92)	n/a
C. rosaceana	2 (A95, A96)	2 (PM76, PM77)

[i]Number of individuals.
[ii](Specimen identifiers) codes correspond to specimen identifiers in Table A1.

Table 2-3
Descriptive Genotyping by Sequencing results from ApeKI and PstI-MspI enzyme digests using the UNEAK pipeline (no reference genome).

	ApeKI		PstI-MspI	
	2[i]	5	2	5
# specimens	102	102	144	144
# unique sequences total	31,860,751	31,860,751	48,521,406	48,521,406
# sequences with greater than 2[i] or 5 reads[ii]	6,707,976	3,265,071	10,030,469	3,388,794
# sequence alignments using a 0.03 error tolerance rate	2,641,530	1,186,525	2,029,321	756,095
# haplotypes	608,970	355,318	391,050	191,210
# tag pairs	304,485	177,659	195,525	95,605
# SNPs found total, mnMAF[iii] 0.05	238,032	151,515	120,362	67,911
# SNPs at loci genotyped 10% +	42,262	57,311	16,543	20,149
# SNPs at loci genotyped 75% +	701	779	1,784	1,841

[i]Read depth thresholds 2 and 5 reads per sequence.
[ii]Replicated from Table 1, describing the raw sequence numbers.
[iii]Minimum minor allele frequency, default value of 0.05.

Table 2-4

Alignment results from ApeKI and PstI-MspI Genotyping by Sequencing analysis to the reference *C. fumiferana* genome (SBW_Refcontig_19April2011) using the Burrows Wheeler Alignment tool.

	ApeKI				**PstI-MspI**			
	2[i]	**5**	**10**	**15**	**2**	**5**	**10**	**15**
Specimens	102	102	102	102	144	144	144	144
Unique sequences[ii]	6,707,976	3,265,071	2,021,933	1,473,593	10,030,469	3,388,794	1805,581	1,299,131
Aligned to unique positions	3,596,692	1,636,443	1,049,564	795,806	4,685,265	1,759,090	939,026	665,675
Aligned to multiple positions	830,719	413,466	262,215	194,604	759,886	283,238	159,319	118,058
Unaligned	2,280,565	1,215,162	710,154	483,183	4,585,318	1,346,466	707,236	515,398
Proportion aligned to unique positions	0.54	0.50	0.52	0.54	0.47	0.52	0.52	0.51

[i]Read depth thresholds 2, 5, 10, and 15.

[ii]Replicated from Table 1, describing the raw sequence numbers.

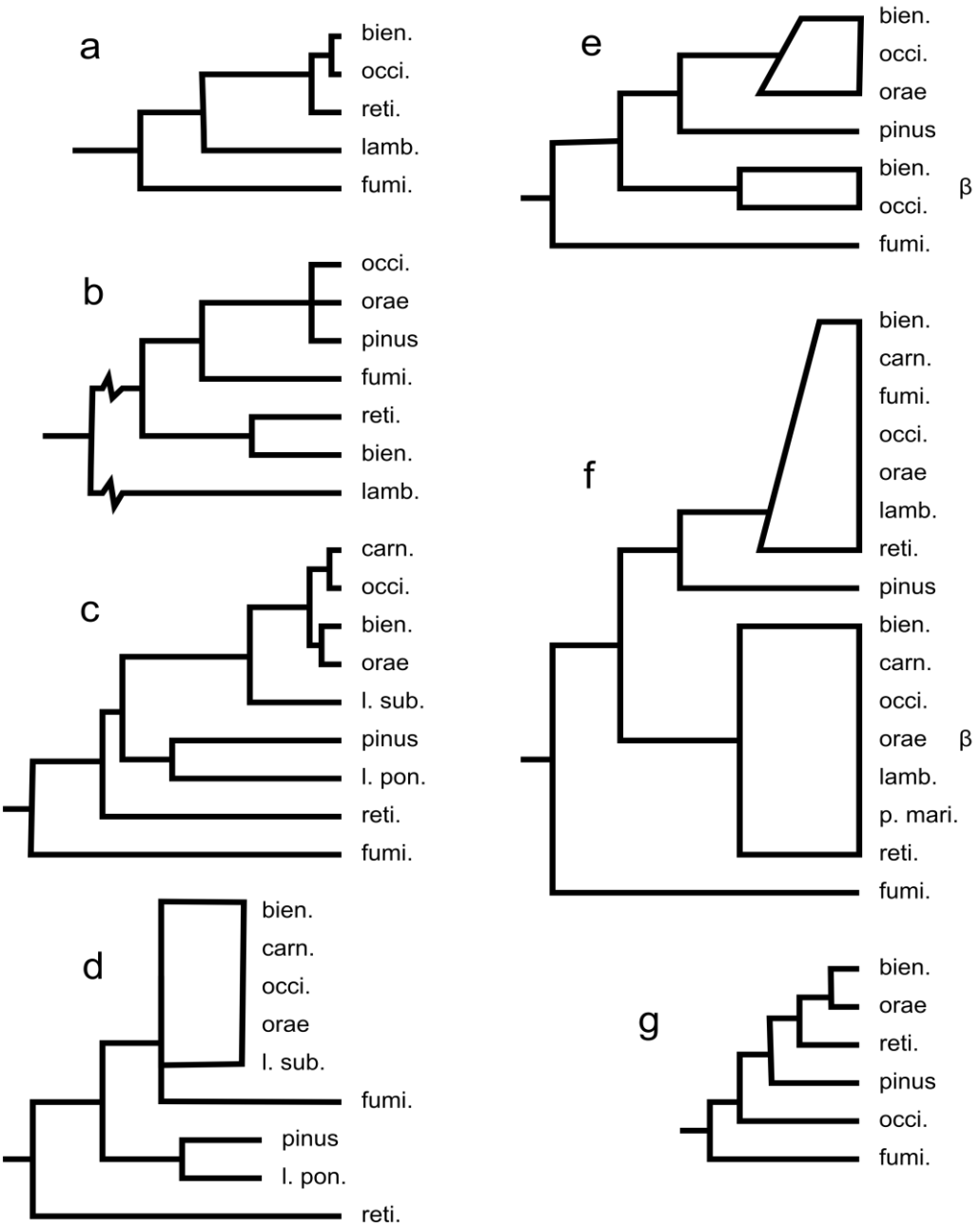


Figure 2-1. Previous phylogenies of the spruce budworm complex: (a) Stock and Castrovillo (1981) from 18 allozyme loci, (b) Castrovillo (1982) from 21 allozyme loci, (c, d) Harvey (1996) from 9 and 12 allozyme loci respectively, (e) Sperling and Hickey (1994, 1995) from mitochondrial COI and COII, (f) Lumley and Sperling (2011a) from mitochondrial COI and COII, (g) and Dombroskie (2011) from mitochondrial COI and part of the nuclear 28S gene. Beta haplotypes (β) are genetically distinct from other haplotypes found in the same species. Boxes indicate unresolved groups, and species names are abbreviated: bien. = *C. biennis*, carn. = *C. carnana*, fumi. = *C. fumiferana*, occi., = *C. occidentalis*, lamb. = *C. lambertiana*, l. sub. = *C. l. subretiniana*, l. pon. = *C. l. ponderosana*, reti. = *C. retiniana*.

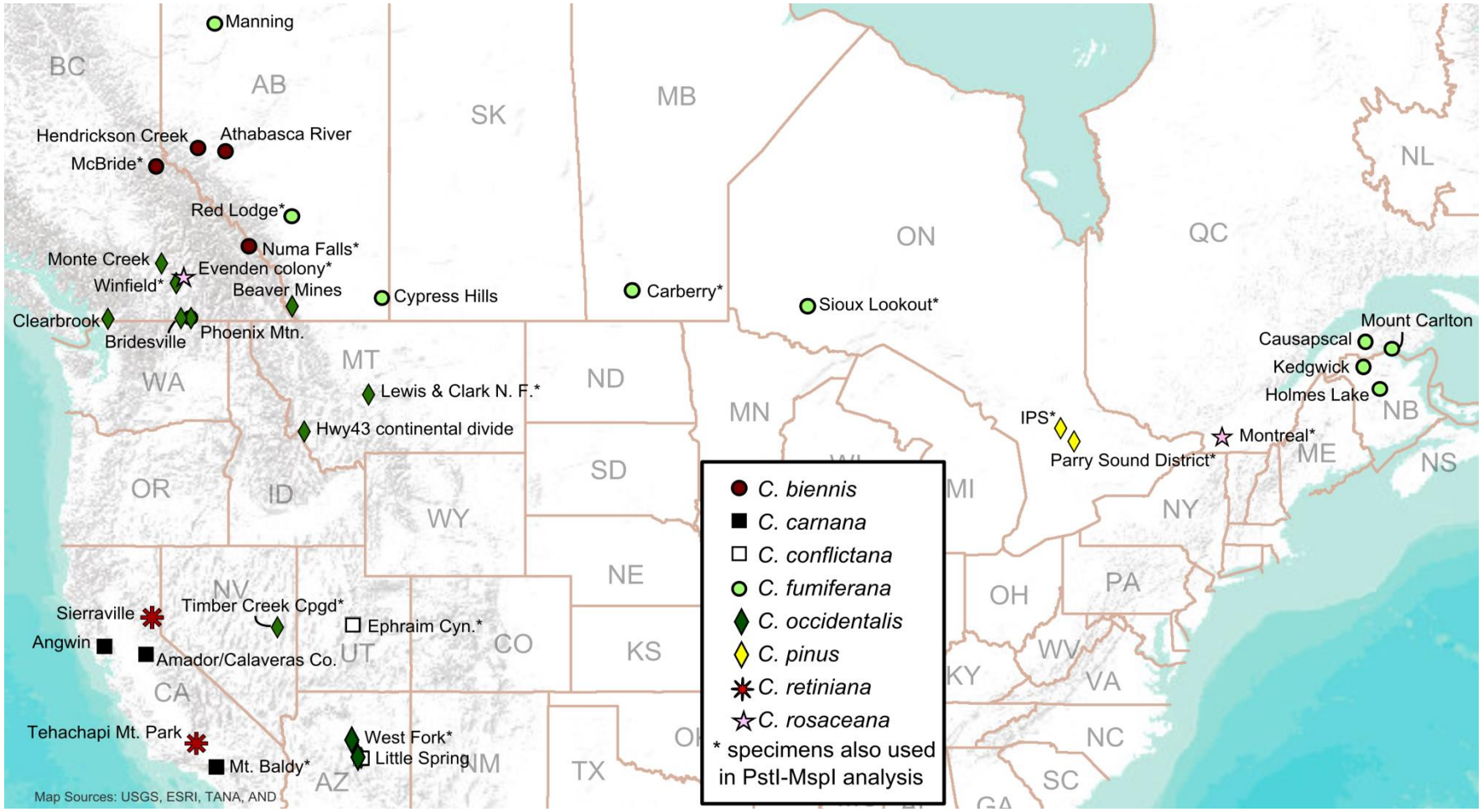


Figure 2-2. Collection locations in North America for eight *Choristoneura* species genotyped using Genotypingn by Sequencing ApeKI analysis.

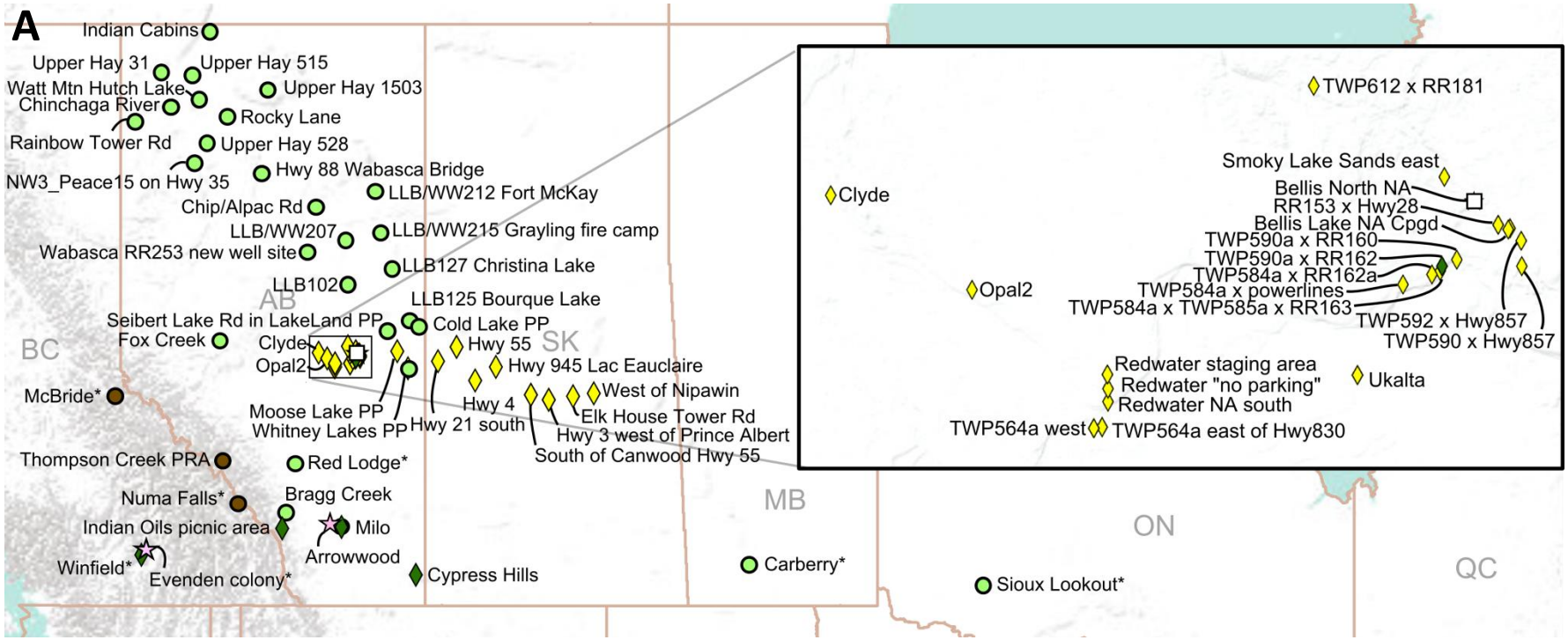


Figure 2-3. Collection locations in North America for seven *Choristoneura* species genotyped using Genotyping by Sequencing PstI-MspI analysis. **A.** Western Canada. See following page for species key.

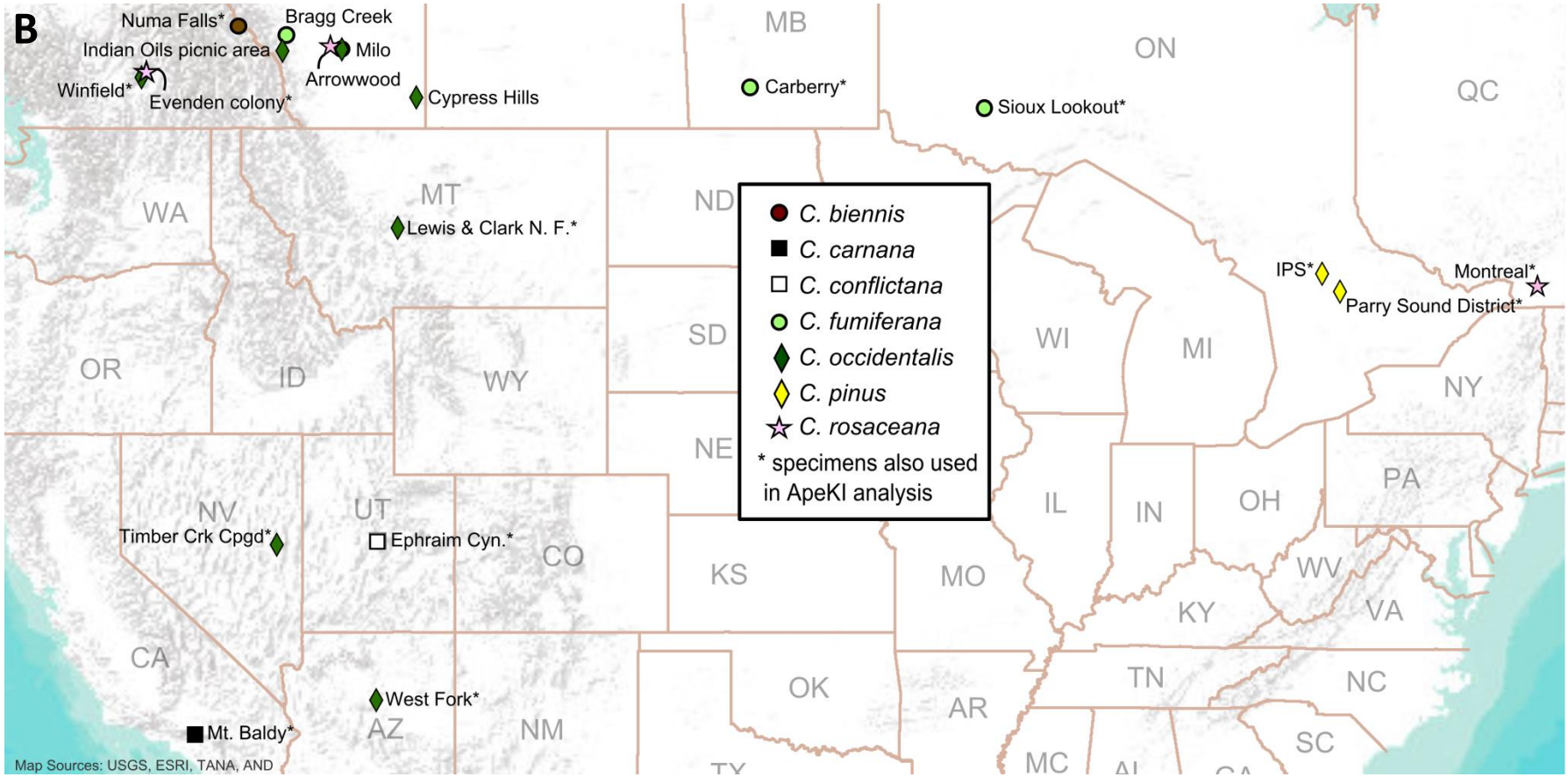


Figure 2-3. cont. Collection locations in North America for seven *Choristoneura* species genotyped using Genotyping by Sequencing PstI-MspI analysis. **B.** USA and eastern Canada. See previous page for northern half of map.

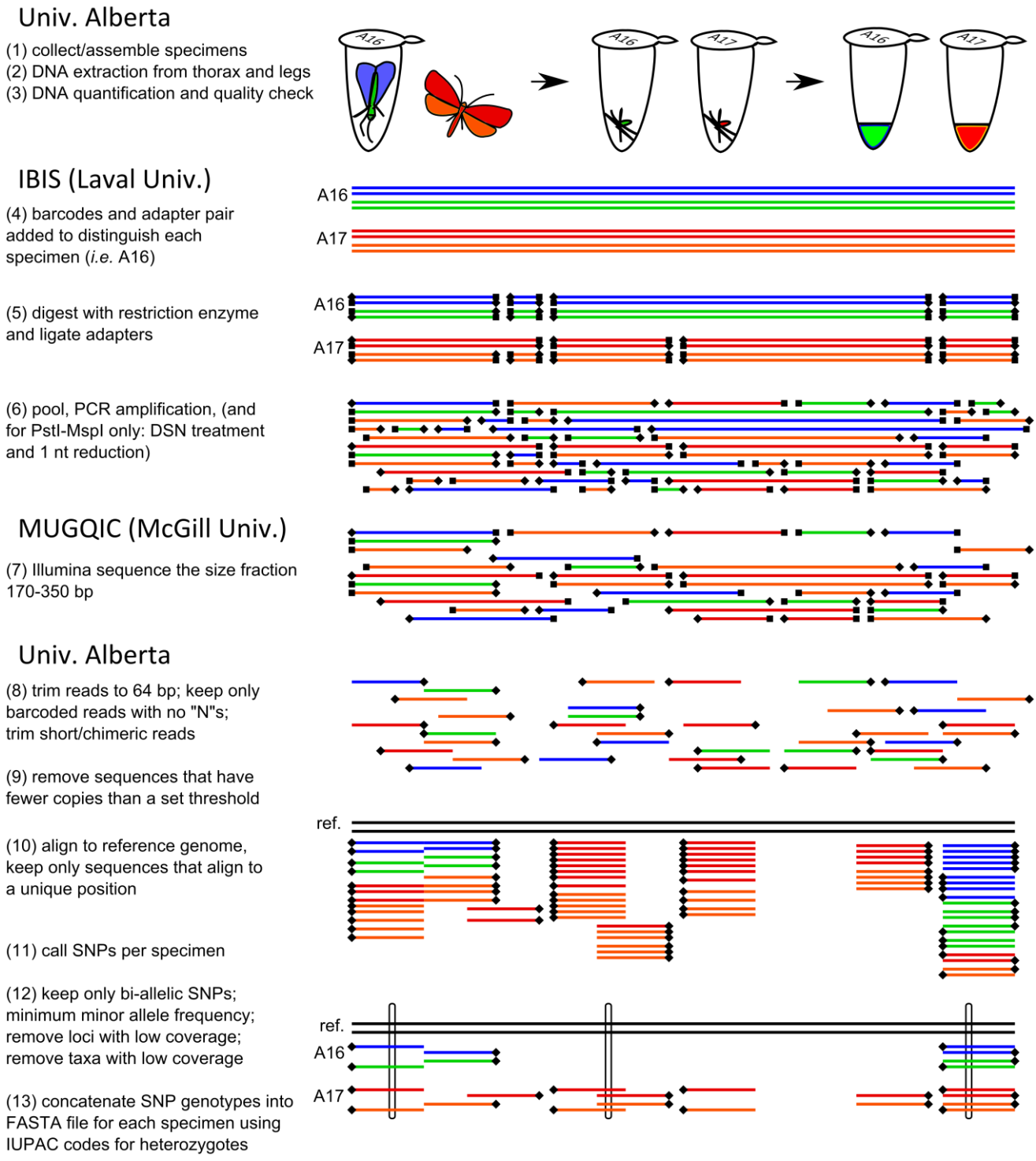


Figure 2-4. Genotyping by sequencing data flow. Specimen selection and DNA preparation (1-3), library preparation (4-6), sequencing (7), and bioinformatics analysis with TASSEL and Burrows Wheeler Alignment (8-13).

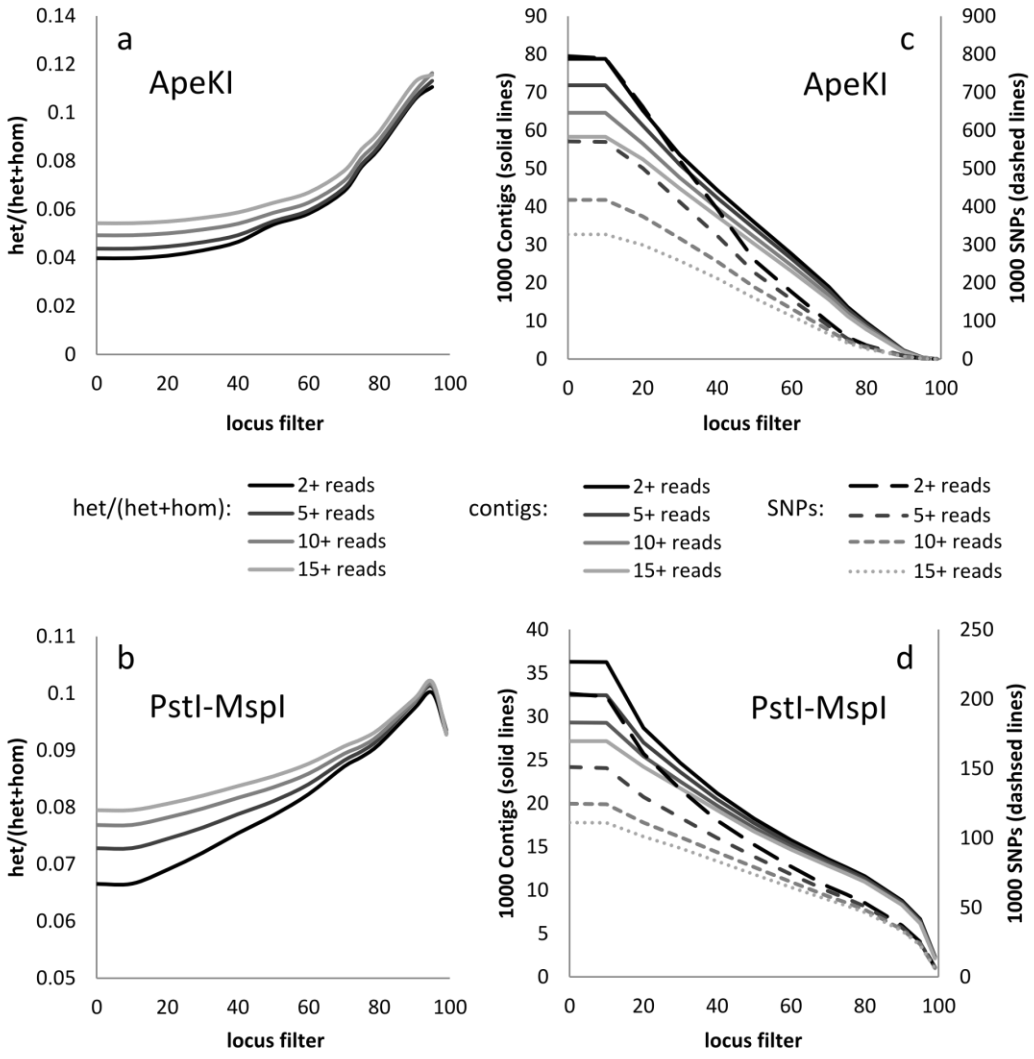


Figure 2-5. Correspondence between the locus filter (which removes loci with less than the minimum proportion of specimens genotyped) and the proportions of heterozygous base calls, contigs, and SNPs. Proportion of heterozygous base calls in (a) ApeKI and (b) PstI-MspI Genotyping by Sequencing (GBS) analysis for read depth thresholds of 2, 5, 10, and 15 reads per sequences. Number of contigs and SNPs in (c) ApeKI and (d) PstI-MspI GBS analysis for read depth thresholds of 2, 5, 10, and 15. The number of contigs is represented by solid lines, and the number of SNPs by dashed lines.

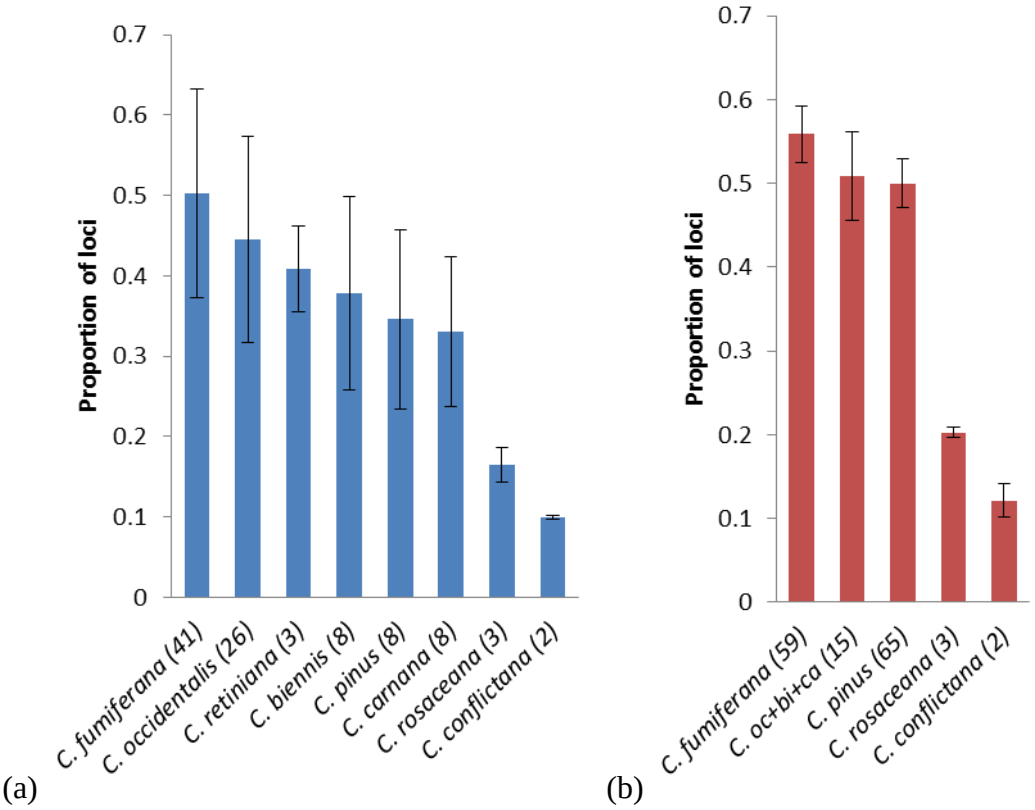


Figure 2-6. Average proportion of loci genotyped for each species. (a) ApeKI dataset (overall average 0.415), (b) and the PstI-MspI dataset (overall average 0.514), at a read depth threshold of 2 and locus genotype coverage threshold of 10%, with standard deviation error bars, and the number of specimens for each species in brackets.

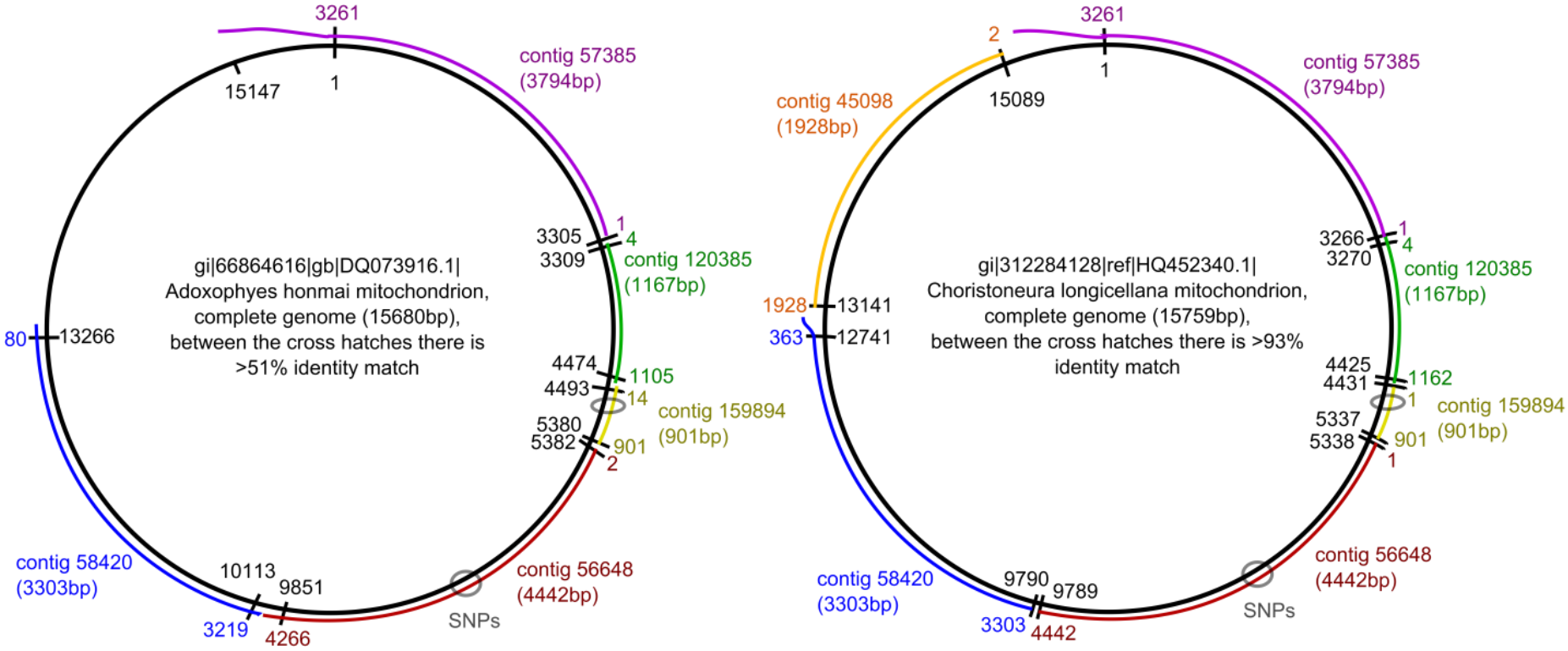


Figure 2-7. Map of *C. fumiferana* reference genome contigs on *A. honmai* and *C. longicellana* mitochondrial genomes. Grey circles indicate the PstI-MspI SNPs found between the ATP6 and COX3 genes (contig 159894) and the ApeKI SNPs found in the NADH5 gene (contig 56648).

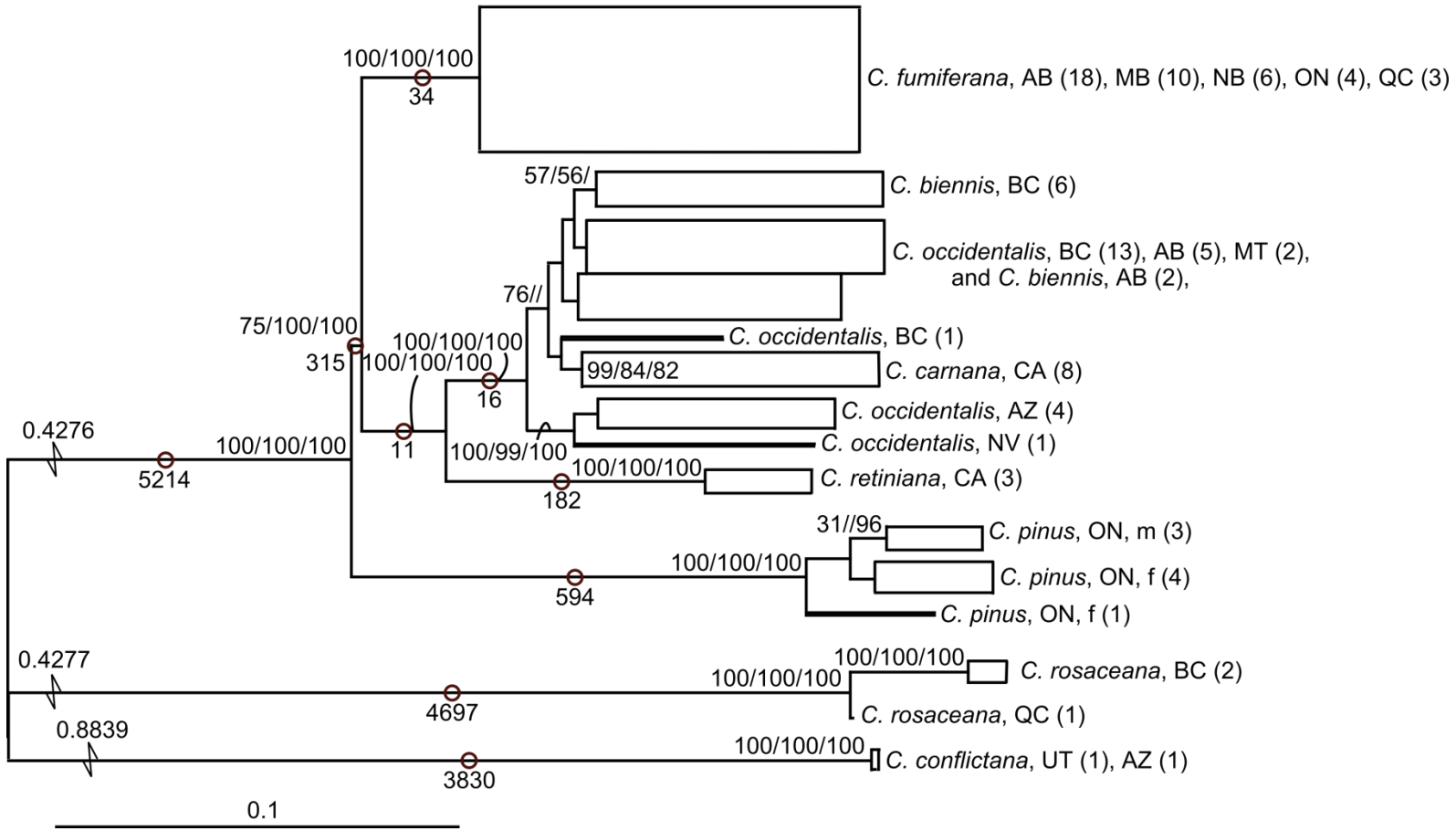


Figure 2-8. ApeKI Maximum Likelihood phylogeny with Maximum Likelihood / Maximum Parsimony / Neighbour Joining bootstrap values and number of apomorphic loci. Numbers of diagnostic or apomorphic loci are indicated for each clade below the circles on the branches. Location, sex (m = male, f = female), and number of specimens are indicated after the species names.

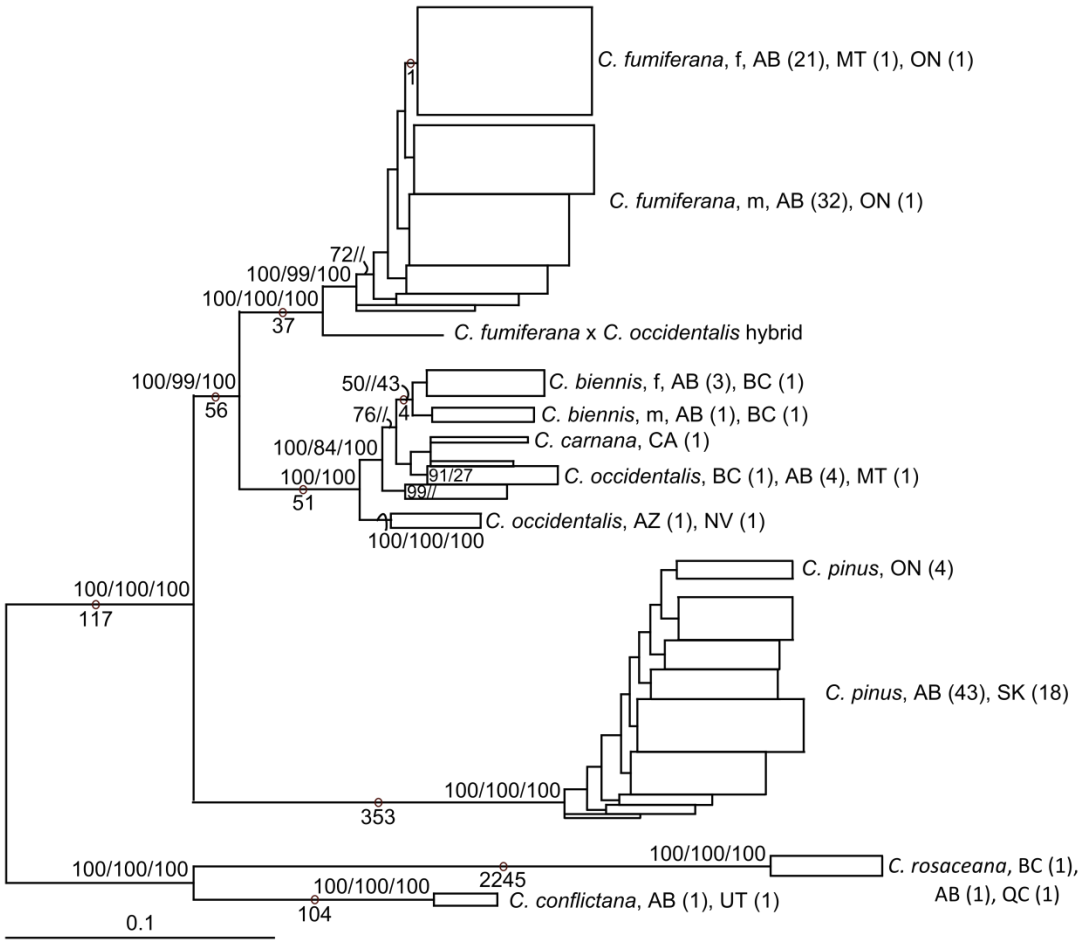


Figure 2-9. PstI-MspI Maximum Likelihood phylogeny with Maximum Likelihood / Maximum Parsimony / Neighbour Joining bootstrap values and number of diagnostic or apomorphic loci. Location, sex (m = male, f = female), and number of specimens are indicated after the species names. The specimen labelled as a hybrid was a genetic intermediate between *C. fumiferana* and *C. occidentalis*.

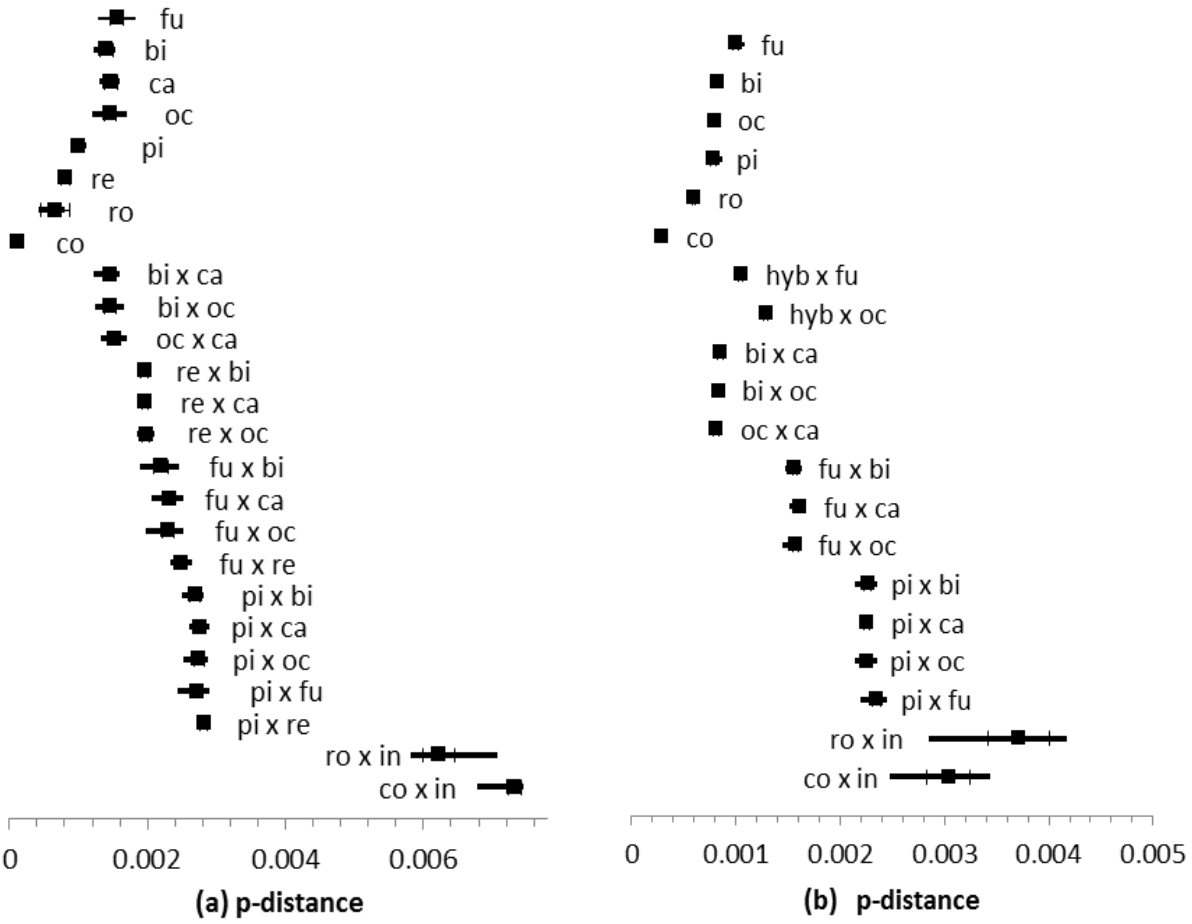


Figure 2-10. Intra- and inter-specific pairwise differences. (a) ApeKI N=99, SNPs=789,600; (b) and PstI-MspI N=144, SNPs=201,748, using the read depth threshold of 2 and locus genotype filter of 10%, p-distance calculated in MEGA5.1, and corrected to include invariant loci. Boxes indicate the mean pairwise distance, thick lines indicate the range, and the thin lines with crossbars indicate the standard deviation. Species names are abbreviated: fu = *C. fumiferana*, bi = *C. biennis*, ca = *C. carnana*, oc = *C. occidentalis*, pi = *C. pinus*, re = *C. retiniana*, ro = *C. rosaceana*, co = *C. conflictana*, hyb = hybrid specimen, and in = ingroup.

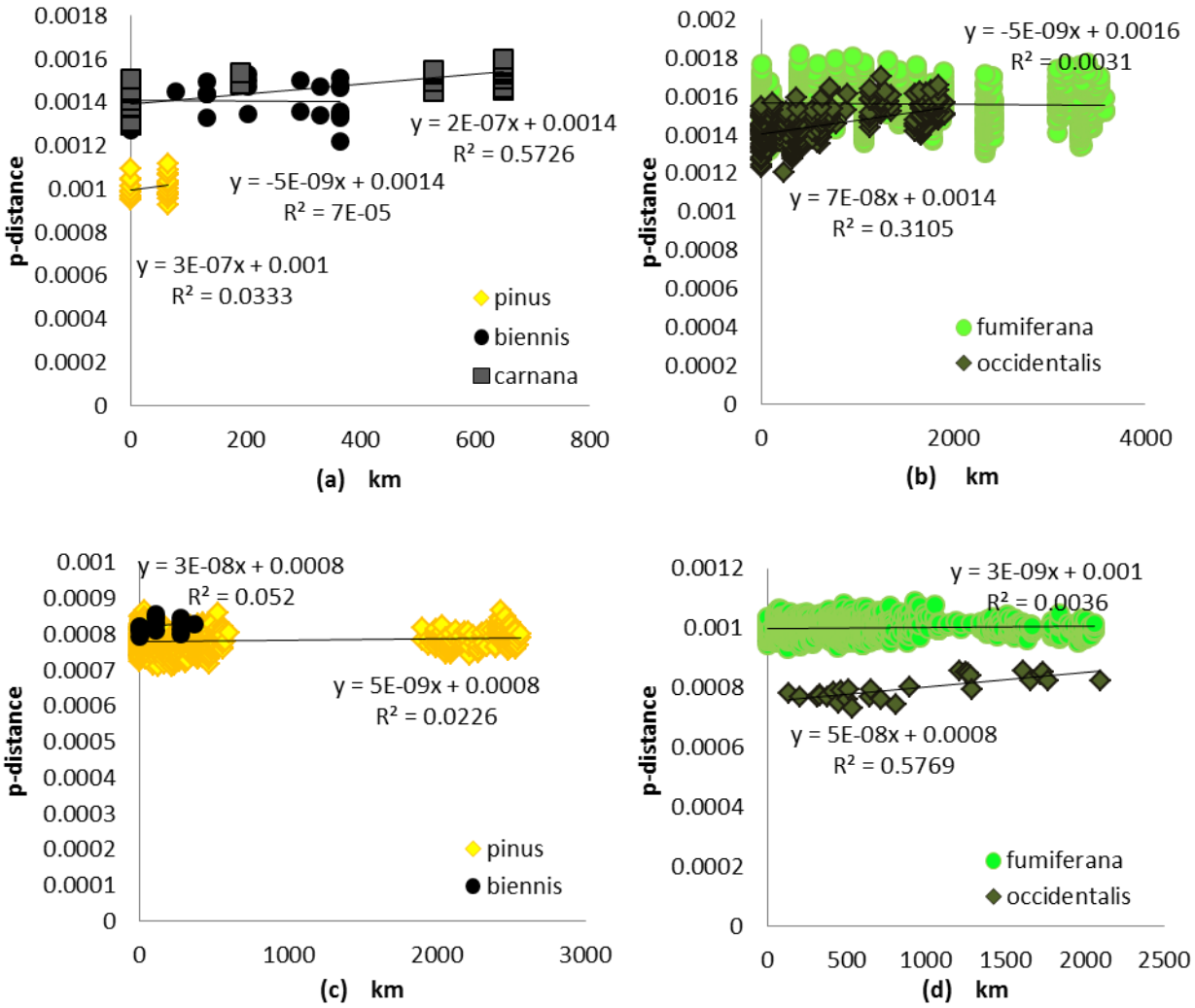


Figure 2-11. Isolation by distance. Linear regression of corrected p-distance and km pairwise comparisons of specimens from the Genotyping by Sequencing (GBS) ApeKI read depth threshold of 2 and locus filter level 10% analysis (a) *C. pinus* (yellow diamond), *C. biennis* (black circle), and *C. carnana* (black square), (b) *C. fumiferana* (green circle) and *C. occidentalis* (dark green diamond), (c, d) and the same species except for *C. carnana* with the GBS PstI-MspI read depth threshold of 2 and locus filter level 10% analysis. The locus filter removes loci with less than the minimum proportion of specimens genotyped.

Appendix A

Table A-1
Specimens and collection information.

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. biennis	CAN: AB: Hendrickson Creek	53.782	-118.351	5.vii-5.xi.2005	Pherom. trap in *Picea glauca*	ASRD	m	A1	2746
C. biennis	CAN: AB: Athabasca River	53.703	-117.157	5.vii-5.xi.2005	Pherom. trap in *P. glauca*	ASRD	m	A2	2757
C. biennis	CAN: BC: McBride	53.300	-120.167	17-18.vi.1992	*Abies lasiocarpa*	N. Humphreys	m	A3	6853
C. biennis	CAN: BC: McBride	53.300	-120.167	17-18.vi.1992	*A. lasiocarpa*	N. Humphreys	m	A4	6855
C. biennis	CAN: BC: McBride	53.300	-120.167	17-18.vi.1992	*A. lasiocarpa*	N. Humphreys	f	A5, PM66	6856
C. biennis	CAN: BC: McBride	53.300	-120.167	25.vi.1992	*A. lasiocarpa*	N. Humphreys	f?	A6	6857
C. biennis	CAN: BC: Numa Falls	51.131	-116.126	26.vi.1992	*Picea engelmannii*	F. Sperling, J. Volney, J. Weber	m	A7	6861
C. biennis	CAN: BC: Numa Falls	51.131	-116.126	26.vi.1992	*P. engelmannii*	F. Sperling, J. Volney, J. Weber	m	A8, PM67	6862
C. biennis	CAN: AB: Thompson Creek PRA	52.012	-116.628	20.vi.2012	*P. glauca*	H. Bird	f	PM137	7776
C. biennis	CAN: AB: Thompson Creek PRA	52.012	-116.628	20.vi.2012	*P. glauca*	H. Bird	f	PM83	7679
C. biennis	CAN: AB: Thompson Creek PRA	52.012	-116.628	20.vi.2012	*P. glauca*	H. Bird	m	PM84	7682
C. biennis	CAN: AB: Thompson Creek PRA	52.012	-116.628	20.vi.2012	*P. glauca*	H. Bird	f	PM93	7714
C. carnana	USA: CA: Mt. Baldy (4 mi N), San Bernardino Co.	34.262	-117.551	17.vii.1995	UV light	J. Powell, F. Sperling	-	A10	823
C. carnana	USA: CA: Angwin, Napa Co.	38.578	-122.448	20.v.1995	*Pseudotsuga menziesii*	F. Sperling	-	A11	824
C. carnana	USA: CA: Mt. Baldy (4 mi N), San Bernardino Co.	34.262	-117.551	17.vii.1995	UV light	J. Powell, F. Sperling	-	A12	906

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. carnana	USA: CA: Amador/Calaveras Co.	38.481	-120.266	26.vii.1998	UV light	D. Rubinoff	-	A13	986
C. carnana	USA: CA: Mt. Baldy (4 mi N), San Bernardino Co.	34.262	-117.551	17.vii.1995	UV light	J. Powell, F. Sperling	m	A14	6878
C. carnana	USA: CA: Mt. Baldy (4 mi N), San Bernardino Co.	34.262	-117.551	17.vii.1995	UV light	J. Powell, F. Sperling	m	A15	6879
C. carnana	USA: CA: Mt. Baldy (4 mi N), San Bernardino Co.	34.262	-117.551	17.vii.1995	UV light	J. Powell, F. Sperling	m	A16, PM68	6880
C. carnana	USA: CA: Angwin, Napa Co.	38.578	-122.448	20.v.1995	*P. menziesii*	F. Sperling	-	A9	800
C. conflictana	USA: AZ: Little Spring, 12 mi N of Flagstaff, Coconino Co.	34.599	-111.228	15.vii.1995	UV light	J. Powell, F. Sperling	-	A17	810
C. conflictana	USA: UT: Ephraim Cyn.-10 mi E of Ephraim, Sanpete Co.	39.360	-111.584	19-20.vii.1996	UV light	J. Powell, F. Sperling	f?	A18, PM69	6881
C. conflictana	CAN: AB: Bellis North NA	54.140	-112.224	24-25.vi.2012	UV light	H. Bird	m	PM75	7772
C. fumiferana	CAN: QC: Causapscal, Metapedia Valley	48.372	-67.232	3.vii.1991	*P. glauca*	FIDS, C. Hébert	f	A19	51
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	-	A20	57
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	-	A21	58
C. fumiferana	CAN: NB: Kedgwick, Hornes Gulch Airstrip	47.633	-67.333	19.vi.1991	*A. balsamea*	FIDS, O.A.M., E. Hurley	m	A22	62
C. fumiferana	CAN: NB: Kedgwick, Hornes Gulch Airstrip	47.633	-67.333	19.vi.1991	*A. balsamea*	FIDS, O.A.M., E. Hurley	m?	A23	63
C. fumiferana	CAN: NB: Holmes Lake	46.950	-66.617	17.vi.1991	*A. balsamea*	FIDS, S. Cormier	-	A24	67
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	f	A25, PM78	72
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	f	A26	73
C. fumiferana	CAN: AB: Manning, Hawk Hills	56.915	-117.609	20.vi.1991	*P. glauca*	FIDS	f	A27	75

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. fumiferana	CAN: AB: Manning, Hawk Hills	56.915	-117.609	20.vi.1991	*P. glauca*	FIDS	-	A28	76
C. fumiferana	CAN: QC: Causapscal, Metapedia Valley	48.372	-67.232	3.vii.1991	*P. glauca*	FIDS, C. Hébert	m	A29	98
C. fumiferana	CAN: QC: Causapscal, Metapedia Valley	48.372	-67.232	3.vii.1991	*P. glauca*	FIDS, C. Hébert	-	A30	99
C. fumiferana	CAN: NB: Kedgwick, Hornes Gulch Airstrip	47.633	-67.333	19.vi.1991	*A. balsamea*	FIDS, O.A.M., E. Hurley	-	A31	108
C. fumiferana	CAN: NB: Kedgwick, Hornes Gulch Airstrip	47.633	-67.333	19.vi.1991	*A. balsamea*	FIDS, O.A.M., E. Hurley	-	A32	110
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A33	159
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A34, PM79	160
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A35	162
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A36	163
C. fumiferana	CAN: AB: Cypress Hills, Battle Creek Cpgd	49.651	-110.029	11.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A37	165
C. fumiferana	CAN: AB: Cypress Hills, Battle Creek Cpgd	49.651	-110.029	11.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A38	166
C. fumiferana	CAN: AB: Cypress Hills, Battle Creek Cpgd	49.651	-110.029	11.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A39	167
C. fumiferana	CAN: AB: Cypress Hills, Battle Creek Cpgd	49.651	-110.029	11.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A40	279
C. fumiferana	CAN: AB: Cypress Hills, Battle Creek Cpgd	49.651	-110.029	11.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A41	280
C. fumiferana	CAN: AB: Cypress Hills, Battle Creek Cpgd	49.651	-110.029	11.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A42	281

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A43	282
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A44	284
C. fumiferana	CAN: AB: Manning, Hawk Hills	56.915	-117.609	20.vi.1991	*P. glauca*	FIDS	-	A45	371
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A46	372
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A47	374
C. fumiferana	CAN: AB: Red Lodge PP	51.943	-114.240	12.vi.1992	*P. glauca*	F. Sperling, J. Volney, J. Weber	-	A48	375
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	-	A50	377
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	-	A51	378
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	-	A52	379
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	-	A53	380
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	-	A54	381
C. fumiferana	CAN: ON: Dewan Township, Ignace District, Sioux Lookout	49.417	-91.667	14.vi.1991	*P. glauca*	FIDS, R. Sajan	-	A55	382
C. fumiferana	CAN: ON: Dewan Township, Ignace District, Sioux Lookout	49.417	-91.667	14.vi.1991	*P. glauca*	FIDS, R. Sajan	-	A56, PM80	383
C. fumiferana	CAN: ON: Dewan Township, Ignace District, Sioux Lookout	49.417	-91.667	14.vi.1991	*P. glauca*	FIDS, R. Sajan	-	A57	385
C. fumiferana	CAN: NB: Mount Carlton	48.167	-66.100	20.vi.1991	*P. glauca*	FIDS, O.A.M., E. Hurley	-	A58	393

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. fumiferana	CAN: AB: Manning, Hawk Hills	56.915	-117.609	20.vi.1991	*P. glauca*	FIDS	-	A59	405
C. fumiferana	CAN: MB: Carberry	49.869	-99.359	17.vi.1991	*P. glauca*	FIDS, M. Grandmaison	f?	A60	6801
C. fumiferana	CAN: ON: Dewan Township, Ignace District, Sioux Lookout	49.417	-91.667	14.vi.1991	*P. glauca*	FIDS, R. Sajan	m	A61, PM81	6806
C. fumiferana	CAN: AB: LLB125 Bourque Lake	54.743	-110.505	31.v.2012	No record	ASRD, H. Bird, B. Brunet, G. Fagua	f	PM100	7729
C. fumiferana	CAN: AB: LLB125 Bourque Lake	54.743	-110.505	31.v.2012	No record	ASRD, H. Bird, B. Brunet, G. Fagua	m	PM101	7730
C. fumiferana	CAN: AB: Hwy 88 Wabasca Bridge	57.444	-115.360	15.vi.2012	*P. glauca*	ASRD	f	PM102	7731
C. fumiferana	CAN: AB: Hwy 88 Wabasca Bridge	57.444	-115.360	15.vi.2012	*P. glauca*	ASRD	m	PM103	7733
C. fumiferana	CAN: AB: Hwy 88 Wabasca Bridge	57.444	-115.360	15.vi.2012	*P. glauca*	ASRD	f	PM104	7734
C. fumiferana	CAN: AB: Hwy 88 Wabasca Bridge	57.444	-115.360	15.vi.2012	*P. glauca*	ASRD	m	PM105	7735
C. fumiferana	CAN: AB: Hwy 88 Wabasca Bridge	57.444	-115.360	15.vi.2012	*P. glauca*	ASRD	m	PM106	7736
C. fumiferana	CAN: AB: Chinchaga River (519)	58.595	-118.332	13.vi.2012	*P. glauca*	ASRD	f	PM107	7737
C. fumiferana	CAN: AB: Chinchaga River (519)	58.595	-118.332	13.vi.2012	*P. glauca*	ASRD	m	PM108	7738
C. fumiferana	CAN: AB: Chinchaga River (519)	58.595	-118.332	13.vi.2012	*P. glauca*	ASRD	f	PM109	7740
C. fumiferana	CAN: AB: Chinchaga River (519)	58.595	-118.332	13.vi.2012	*P. glauca*	ASRD	m	PM110	7742
C. fumiferana	CAN: AB: Indian Cabins	59.862	-117.037	14.vi.2012	*P. glauca*	ASRD	m	PM111	7744
C. fumiferana	CAN: AB: Indian Cabins	59.862	-117.037	14.vi.2012	*P. glauca*	ASRD	f	PM112	7745
C. fumiferana	CAN: AB: Indian Cabins	59.862	-117.037	14.vi.2012	*P. glauca*	ASRD	m	PM113	7746
C. fumiferana	CAN: AB: Rainbow Tower Rd, S of Rainbow Lake (530)	58.350	-119.509	13.vi.2012	*P. glauca*	ASRD	f	PM114	7747
C. fumiferana	CAN: AB: Upper Hay Site 1503	58.892	-115.146	15.vi.2012	*P. glauca*	ASRD	f	PM115	7749
C. fumiferana	CAN: AB: Upper Hay Site 1503	58.892	-115.146	15.vi.2012	*P. glauca*	ASRD	f	PM116	7750
C. fumiferana	CAN: AB: Upper Hay Site 1503	58.892	-115.146	15.vi.2012	*P. glauca*	ASRD	f	PM117	7752

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. fumiferana	CAN: AB: Upper Hay 31	59.189	-118.641	13.vi.2012	*P. glauca*	ASRD	m	PM118	7753
C. fumiferana	CAN: AB: Upper Hay 515	59.144	-117.637	13.vi.2012	*P. glauca*	ASRD	m	PM119	7755
C. fumiferana	CAN: AB: Watt Mountain, Hutch Lake	58.726	-117.386	14.vi.2012	*P. glauca*	ASRD	m	PM120	7756
C. fumiferana	CAN: AB: Watt Mountain, Hutch Lake	58.726	-117.386	14.vi.2012	*P. glauca*	ASRD	f	PM121	7758
C. fumiferana	CAN: AB: new well site N of Wabaska on RR253	56.030	-113.860	28.v.2012	*P. glauca*	ASRD, H. Bird, B. Brunet, G. Fagua	f	PM122	7759
C. fumiferana	CAN: AB: new well site N of Wabaska on RR253	56.030	-113.860	28.v.2012	*P. glauca*	ASRD, H. Bird, B. Brunet, G. Fagua	f	PM123	7760
C. fumiferana	CAN: AB: LLB102	55.430	-112.513	1.vi.2012	*P. glauca*	ASRD	f	PM124	7761
C. fumiferana	CAN: AB: LLB102	55.430	-112.513	1.vi.2012	*P. glauca*	ASRD	m	PM125	7762
C. fumiferana	CAN: AB: LLB102	55.430	-112.513	1.vi.2012	*P. glauca*	ASRD	f	PM126	7763
C. fumiferana	CAN: AB: LLB127 Christina Lake	55.719	-111.048	30.v.2012	*A. balsamea*	ASRD, H. Bird, B. Brunet, G. Fagua	m	PM127	7764
C. fumiferana	CAN: AB: LLB/WW207	56.244	-112.581	1.vi.2012	*P. glauca*	ASRD	f	PM128	7765
C. fumiferana	CAN: AB: LLB/WW212 Fort McKay	57.121	-111.630	5.vi.2012	*P. glauca*	ASRD	m	PM129	7766
C. fumiferana	CAN: AB: LLB/WW215 Grayling fire camp	56.379	-111.444	5.vi.2012	*P. glauca*	ASRD	m	PM130	7767
C. fumiferana	CAN: AB: LLB/WW215 Grayling fire camp	56.379	-111.444	5.vi.2012	*P. glauca*	ASRD	m	PM131	7768
C. fumiferana	CAN: AB: LLB/WW215 Grayling fire camp	56.379	-111.444	5.vi.2012	*P. glauca*	ASRD	m	PM132	7769
C. fumiferana	CAN: AB: Rocky lane	58.442	-116.483	15.vi.2012	*P. glauca*	ASRD	m	PM133	7770
C. fumiferana	CAN: AB: Hwy 88 Wabasca Bridge	57.444	-115.360	15.vi.2012	*P. glauca*	ASRD	m	PM134	7771
C. fumiferana	CAN: AB: NW3_Peace15 on Hwy 35	57.624	-117.554	6.vi.2012	*P. glauca*	ASRD, H. Bird, B. Brunet, G. Fagua	f	PM135	7773

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. fumiferana	CAN: AB: Upper Hay528	57.975	-117.132	6.vi.2012	*P. glauca*	ASRD, H. Bird, B. Brunet, G. Fagua	m	PM138	7777
C. fumiferana	CAN: AB: SW4_Woodland near Fox Creek	54.367	-116.719	4.vi.2012	*P. glauca*	ASRD, H. Bird, B. Brunet, G. Fagua	m	PM139	7778
C. fumiferana	CAN: AB: Whitney Lakes PP W of RR43	53.833	-110.526	10-11.vii.2012	*C. fumiferana* lure	H. Bird	m	PM140	7779
C. fumiferana	CAN: AB: Whitney Lakes PP W of RR43	53.833	-110.526	10-11.vii.2012	*C. fumiferana* lure	H. Bird	m	PM141	7780
C. fumiferana	CAN: AB: Cold Lake PP N shore	54.640	-110.180	12-13.vii.2012	*C. fumiferana* lure	H. Bird	m	PM142	7782
C. fumiferana	CAN: AB: Seibert Lake Rd in Lakeland PP, at pull out	54.552	-111.229	13-14.vii.2012	*C. fumiferana* lure	H. Bird	m	PM143	7783
C. fumiferana	CAN: AB: Seibert Lake Rd in Lakeland PP, at pull out	54.552	-111.229	13-14.vii.2012	*C. fumiferana* lure	H. Bird	m	PM144	7784
C. fumiferana	CAN: AB: Milo, Godkin homestead	50.664	-112.759	16.vi.2012	*Picea pungens*	H. Bird, C. & C. & C. Squire	m	PM85	7685
C. fumiferana	CAN: AB: Bragg Creek, W of river	50.955	-114.566	18.vi.2012	*P. glauca*	H. Bird, B. Leibel	m	PM86	7687
C. fumiferana	CAN: AB: Chip/Alpac Rd T-90 R23	56.846	-113.580	29.v.2012	*P. glauca*	ASRD, H. Bird, B. Brunet, G. Fagua	f	PM88	7693
C. fumiferana	CAN: AB: Chip/Alpac Rd T-90 R23	56.846	-113.580	29.v.2012	*P. glauca*	ASRD, H. Bird, B. Brunet, G. Fagua	m	PM89	7694
C. fumiferana	CAN: AB: Chinchaga River (519)	58.595	-118.332	13.vi.2012	*P. glauca*	ASRD	m	PM90	7698
C. fumiferana	CAN: AB: Chinchaga River (519)	58.595	-118.332	13.vi.2012	*P. glauca*	ASRD	m	PM91	7699
C. fumiferana	CAN: AB: Chinchaga River (519)	58.595	-118.332	13.vi.2012	*P. glauca*	ASRD	f	PM92	7700
C. fumiferana	CAN: AB: Upper Hay Site 1503	58.892	-115.146	15.vi.2012	*P. glauca*	ASRD	m	PM94	7718
C. fumiferana	CAN: AB: Upper Hay 515	59.144	-117.637	13.vi.2012	*P. glauca*	ASRD	f	PM95	7720
C. fumiferana	CAN: AB: Rainbow Tower Rd, S of Rainbow Lake (530)	58.350	-119.509	13.vi.2012	*P. glauca*	ASRD	f	PM96	7721

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. fumiferana	CAN: AB: Watt Mountain, Hutch Lake	58.726	-117.386	14.vi.2012	*P. glauca*	ASRD	m	PM97	7724
C. fumiferana	CAN: AB: Indian Cabins	59.862	-117.037	14.vi.2012	*P. glauca*	ASRD	m	PM99	7726
C. n.sp.#3* nr. *occidentalis	CAN: BC: Clearbrook	49.054	-122.306	early spring, 1993	*P. menziesii* (overwintering)	FIDS, T. Gray	-	A62	419
C. n.sp.#3* nr. *occidentalis	CAN: BC: Clearbrook	49.054	-122.306	early spring, 1993	*P. menziesii* (overwintering)	FIDS, T. Gray	-	A63	420
C. occidentalis	CAN: AB: WSBW Beaver Mines; ~8-9 km SW on Hwy 774	49.419	-114.251	6.vi.2009	*P. glauca*	B. Brunet	m	A235	BB2009-152e
C. occidentalis	CAN: AB: WSBW Beaver Mines; ~8-9 km SW on Hwy 774	49.419	-114.251	12-21.vii.2009	*C. fumiferana* lure in *P. glauca*	B. Brunet	m	A236	BB2009-204a.1
C. occidentalis	CAN: AB: WSBW Beaver Mines; ~8-9 km SW on Hwy 774	49.419	-114.251	12-21.vii.2009	*C. fumiferana* lure in *P. glauca*	B. Brunet	m	A237	BB2009-204a.2
C. occidentalis	CAN: AB: WSBW Beaver Mines; ~8-9 km SW on Hwy 774	49.419	-114.251	12-21.vii.2009	*C. fumiferana* lure in *P. glauca*	B. Brunet	m	A238	BB2009-204b.1
C. occidentalis	CAN: AB: WSBW Beaver Mines; ~8-9 km SW on Hwy 774	49.419	-114.251	12-21.vii.2009	*C. fumiferana* lure in *P. glauca*	B. Brunet	m	A239	BB2009-204b.2
C. occidentalis	CAN: BC: Greenwood, Phoenix Mtn.	49.083	-118.683	15.vi.1992	*P. menziesii*	FIDS, A. Stewart	m	A49	376
C. occidentalis	CAN: BC: Winfield	50.080	-119.310	21.vi.1991	*P. menziesii*	FIDS, J. Hodge	f	A64	47
C. occidentalis	CAN: BC: Winfield	50.080	-119.310	21.vi.1991	*P. menziesii*	FIDS, J. Hodge	-	A65	96
C. occidentalis	CAN: BC: Winfield	50.080	-119.310	21.vi.1991	*P. menziesii*	FIDS, J. Hodge	m	A66	97
C. occidentalis	CAN: BC: Greenwood, Phoenix Mtn.	49.083	-118.683	15.vi.1992	*P. menziesii*	FIDS, A. Stewart	-	A67	188
C. occidentalis	CAN: BC: Monte Creek	50.646	-119.949	20.vi.1991	*P. menziesii*	FIDS, P. Koot	-	A68	351
C. occidentalis	CAN: BC: Winfield	50.080	-119.310	21.vi.1991	*P. menziesii*	FIDS, J. Hodge	-	A69	355
C. occidentalis	CAN: BC: Winfield	50.080	-119.310	21.vi.1991	*P. menziesii*	FIDS, J. Hodge	-	A70	359

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. occidentalis	CAN: BC: Greenwood, Phoenix Mtn.	49.083	-118.683	15.vi.1992	*P. menziesii*	FIDS, A. Stewart	-	A71	360
C. occidentalis	CAN: BC: Greenwood, Phoenix Mtn.	49.083	-118.683	15.vi.1992	*P. menziesii*	FIDS, A. Stewart	-	A72	361
C. occidentalis	CAN: BC: Greenwood, Phoenix Mtn.	49.083	-118.683	15.vi.1992	*P. menziesii*	FIDS, A. Stewart	-	A73	362
C. occidentalis	CAN: BC: Greenwood, Phoenix Mtn.	49.083	-118.683	15.vi.1992	*P. menziesii*	FIDS, A. Stewart	-	A74	363
C. occidentalis	CAN: BC: Bridesville (McKinney Creek)	49.072	-119.115	16.vi.1992	*P. menziesii / Pinus contorta*	FIDS, A. Stewart	-	A75	367
C. occidentalis	USA: AZ: Little Spring, 12 mi N of Flagstaff, Coconino Co.	34.599	-111.228	20.vii.1993	*P. menziesii*	F. Sperling	-	A76	818
C. occidentalis	USA: AZ: Little Spring, 12 mi N of Flagstaff, Coconino Co.	34.599	-111.228	20.vii.1993	*P. menziesii*	F. Sperling	-	A77	821
C. occidentalis	USA: AZ: Little Spring, 12 mi N of Flagstaff, Coconino Co.	34.599	-111.228	20.vii.1993	*P. menziesii*	F. Sperling	-	A78	825
C. occidentalis	USA: MT: Lewis & Clark NF, Little Bent Mtns SE of Great Falls on Hwy 89, nr Moose Creek Cpgd	46.804	-110.912	3.vii.2007	*P. menziesii*	Lep Soc Trip 2007	-	A79, PM70	3620
C. occidentalis	USA: MT: Hwy 43 - ~10-15 mi E of contin'l divide on logging rd	45.653	-113.710	4.vii.2007	*P. menziesii*	Lep Soc Trip 2007	-	A80	3631
C. occidentalis	CAN: BC: Winfield, McKinley Rd	50.080	-119.310	21.vi.1991	*P. menziesii*	FIDS, J. Hodge	f?	A81, PM71	6814
C. occidentalis	USA: NV: Timber Creek Cpgd, 20 mi NE of Ely, White Pine Co.	39.249	-114.878	18.vii.1996	UV light	J. Powell, F. Sperling	m	A82, PM72	6827
C. occidentalis	USA: AZ: West Fork, 16 mi SW of Flagstaff, Coconino Co.	35.189	-111.620	13-16.vii.1995	UV light	J. Powell, F. Sperling	m	A83, PM73	6828
C. occidentalis	CAN: AB: Cypress Hills, Lodgepole Cpgd campsite # 21	49.647	-110.298	16.vi.2012	*P. glauca*	G. Fagua	f	PM136	7774

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. occidentalis	CAN: AB: TWP 590a x RR 162	54.064	-112.288	17-19.vii.2012	*C. fumiferana* lure	H. Bird	m	PM82	7677
C. occidentalis	CAN: AB: Indian Oils picnic area	50.622	-114.697	18.vi.2012	*P. glauca / pungens*	H. Bird, B. Leibel	m	PM87	7692
C. occidentalis	CAN: AB: Milo, Godkin homestead	50.664	-112.759	16.vi.2012	*P. pungens*	H. Bird, C. & C. & C. Squire	m	PM98	7725
C. pinus	CAN: ON: Insect Production Services (IPS), Sault Ste. Marie	45.769	-80.615	29.vi.2011	Collected from Brit, ON	CFS GLFC	m	A240	JPBW.2
C. pinus	CAN: ON: Harrison Township, Parry Sound District	45.354	-80.036	19.vi.1991	*Pinus banksiana*	FIDS	f	A84	6807
C. pinus	CAN: ON: Harrison Township, Parry Sound District	45.354	-80.036	19.vi.1991	*P. banksiana*	FIDS	m	A85, PM62	6808
C. pinus	CAN: ON: Henney Township, Parry Sound District	45.354	-80.036	20.vi.1991	*P. banksiana*	FIDS	m	A86, PM63	6810
C. pinus	CAN: ON: Insect Production Services (IPS), Sault Ste. Marie	45.769	-80.615	29.vi.2011	Collected from Brit, ON	CFS GLFC	f	A87, PM64	6863
C. pinus	CAN: ON: Insect Production Services (IPS), Sault Ste. Marie	45.769	-80.615	29.vi.2011	Collected from Brit, ON	CFS GLFC	f	A88	6864
C. pinus	CAN: ON: Insect Production Services (IPS), Sault Ste. Marie	45.769	-80.615	29.vi.2011	Collected from Brit, ON	CFS GLFC	f	A89	6865
C. pinus	CAN: ON: Insect Production Services (IPS), Sault Ste. Marie	45.769	-80.615	29.vi.2011	Collected from Brit, ON	CFS GLFC	f	A90, PM65	6866
C. pinus	CAN: SK: Hwy 3 W of Prince Albert	53.225	-105.942	27-29.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM1	6884
C. pinus	CAN: AB: TWP 564a E of Hwy 830	53.876	-112.962	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM10	7603
C. pinus	CAN: AB: TWP 564a W of Hwy 830	53.875	-112.978	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM11	7607
C. pinus	CAN: AB: TWP 564a W of Hwy 830	53.875	-112.978	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM12	7608

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. pinus	CAN: AB: TWP 564a W of Hwy 830	53.875	-112.978	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM13	7613
C. pinus	CAN: AB: TWP 564a W of Hwy 830	53.875	-112.978	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM14	7616
C. pinus	CAN: AB: Hwy 855 near Ukalta	53.937	-112.456	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM15	7617
C. pinus	CAN: AB: Hwy 855 near Ukalta	53.937	-112.456	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM16	7619
C. pinus	CAN: AB: Hwy 855 near Ukalta	53.937	-112.456	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM17	7620
C. pinus	CAN: AB: Hwy 855 near Ukalta	53.937	-112.456	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM18	7621
C. pinus	CAN: AB: Hwy 855 near Ukalta	53.937	-112.456	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM19	7622
C. pinus	CAN: SK: Hwy 3 W of Prince Albert	53.225	-105.942	27-29.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM2	6887
C. pinus	CAN: AB: Hwy 855 near Ukalta	53.937	-112.456	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM20	7624
C. pinus	CAN: AB: Hwy 855 near Ukalta	53.937	-112.456	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM21	7626
C. pinus	CAN: AB: Hwy 855 near Ukalta	53.937	-112.456	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM22	7627
C. pinus	CAN: SK: Hwy 3 W of Prince Albert	53.225	-105.942	27-29.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM23	7629
C. pinus	CAN: SK: Hwy 3 W of Prince Albert	53.225	-105.942	27-29.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM24	7631
C. pinus	CAN: SK: Hwy 3 W of Prince Albert	53.225	-105.942	27-29.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM25	7633
C. pinus	CAN: SK: S of Canwood Hwy 55	53.323	-106.525	26-27.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM26	7636

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. pinus	CAN: SK: S of Canwood Hwy 55	53.323	-106.525	26-27.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM27	7638
C. pinus	CAN: SK: S of Canwood Hwy 55	53.323	-106.525	26-27.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM28	7639
C. pinus	CAN: SK: Hwy 55	54.255	-108.979	24-25.vii.2012	UV light trap in *P. banksiana*	H. Bird	m	PM29	7640
C. pinus	CAN: SK: Hwy 55	54.255	-108.979	24-25.vii.2012	UV light trap in *P. banksiana*	H. Bird	m	PM3	6888
C. pinus	CAN: SK: Hwy 4	53.614	-108.362	25-26.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM30	7641
C. pinus	CAN: SK: Hwy 21 S	53.981	-109.592	24-25.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM31	7642
C. pinus	CAN: AB: Hwy 18 x TWP 242	54.146	-113.500	23-24.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM32	7643
C. pinus	CAN: AB: TWP 584 x RR 224	54.037	-113.221	23-24.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM33	7645
C. pinus	CAN: SK: Hwy 945 Lac Eauclaire	53.871	-107.690	25-26.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM34	7647
C. pinus	CAN: AB: Bellis Lake NA Cpgd	54.108	-112.153	14-15.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM35	7651
C. pinus	CAN: AB: Bellis Lake NA Cpgd	54.108	-112.153	14-15.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM36	7652
C. pinus	CAN: AB: TWP 584a x RR 162a	54.058	-112.292	19-20.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM37	7653
C. pinus	CAN: AB: TWP 584a x TWP 585a x RR 163	54.054	-112.308	19-20.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM38	7654
C. pinus	CAN: AB: TWP 584a x powerlines	54.042	-112.365	19-20.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM39	7655
C. pinus	CAN: SK: S of Canwood Hwy 55	53.323	-106.525	26-27.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM4	6891

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. pinus	CAN: AB: TWP 590a x RR 162	54.064	-112.288	17-19.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM40	7656
C. pinus	CAN: AB: TWP 590a x RR 162	54.064	-112.288	17-19.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM41	7657
C. pinus	CAN: AB: TWP 590a x RR 160	54.072	-112.259	17-19.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM42	7659
C. pinus	CAN: AB: TWP 590 x Hwy 857	54.064	-112.130	17-19.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM43	7660
C. pinus	CAN: AB: RR 153 x Hwy 28	54.112	-112.177	20-23.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM44	7661
C. pinus	CAN: AB: RR 153 x Hwy 28	54.112	-112.177	20-23.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM45	7662
C. pinus	CAN: AB: TWP 612 x RR 181	54.273	-112.543	20-23.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM46	7663
C. pinus	CAN: AB: Moose Lake PP Cpgd	54.172	-110.917	12.vii.2012	*P. banksiana*	H. Bird	m	PM47	7665
C. pinus	CAN: AB: Smoky Lake Sands East	54.168	-112.283	9.vii.2012	*P. banksiana*	H. Bird	m	PM48	7668
C. pinus	CAN: SK: Elk House Tower Rd	53.300	-105.151	27-29.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM49	7670
C. pinus	CAN: SK: S of Canwood Hwy 55	53.323	-106.525	26-27.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM5	6892
C. pinus	CAN: SK: W of Nipawin	53.351	-104.468	27-29.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM50	7672
C. pinus	CAN: AB: Whitney Lakes PP, RR45	53.846	-110.575	11.vii.2012	*P. banksiana*	H. Bird	m	PM51	7673
C. pinus	CAN: AB: TWP 592 x Hwy 857	54.094	-112.131	15-17.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM52	7674
C. pinus	CAN: AB: Bellis Lake NA Cpgd	54.106	-112.156	15-17.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM53	7676
C. pinus	CAN: AB: Redwater NA TWP 571a x RR 205	53.906	-112.951	7-13.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM54	7701

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. pinus	CAN: AB: Redwater NA TWP 571a x RR 205	53.906	-112.951	7-13.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM55	7703
C. pinus	CAN: AB: Redwater NA "no parking" sign	53.921	-112.950	7-13.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM56	7705
C. pinus	CAN: AB: Redwater NA "no parking" sign	53.921	-112.950	7-13.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM57	7706
C. pinus	CAN: AB: Redwater NA staging area	53.938	-112.952	7-13.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM58	7707
C. pinus	CAN: AB: Redwater NA staging area	53.938	-112.952	7-13.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM59	7708
C. pinus	CAN: AB: TWP 564a E of Hwy 830	53.876	-112.962	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM6	6896
C. pinus	CAN: SK: Hwy 3 W of Prince Albert	53.225	-105.942	27-29.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM60	7709
C. pinus	CAN: SK: S of Canwood Hwy 55	53.323	-106.525	26-27.vii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM61	7711
C. pinus	CAN: AB: TWP 564a E of Hwy 830	53.876	-112.962	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM7	6897
C. pinus	CAN: AB: TWP 564a E of Hwy 830	53.876	-112.962	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM8	6899
C. pinus	CAN: AB: TWP 564a E of Hwy 830	53.876	-112.962	1-7.viii.2012	*C. pinus* lure in *P. banksiana*	H. Bird	m	PM9	7601
C. retiniana	USA: CA: Tehachapi Mt. Park, Kern Co.	35.135	-118.440	13.vii.1995	Beating	J. Powell, F. Sperling	-	A91	827
C. retiniana	USA: CA: Sierraville, 4 mi SE, Sierra Co.	39.589	-120.366	27-28.vii.1995	No record	J. Powell	-	A92	828
C. retiniana	USA: CA: Sierraville, 4 mi SE, Sierra Co.	39.589	-120.366	27-28.vii.1995	No record	J. Powell	-	A93	829
C. rosaceana	CAN: AB: Evenden lab colony, Univ. of Alberta	50.4	-119.3	24.xi.2010	Okanagan Valley	M. Evenden	f	A94	6867

Species	Sample locality[i]	Lat. (deg)	Long. (deg)	Date collected	Larval host / Coll. method	Collector[ii]	Sex	RE[iii] sample#	DNA #[iv]
C. rosaceana	CAN: AB: Evenden lab colony, Univ. of Alberta	50.4	-119.3	24.xi.2010	Okanagan Valley	M. Evenden	f	A95, PM76	6871
C. rosaceana	CAN: QC: Montreal	45.510	-73.550	15.vii.1992	UV light	J.-F. Landry	m	A96, PM77	6882
C. rosaceana	CAN: AB: Arrowwood, Bird farm	50.751	-113.128	6-7.vii.2012	UV light	H. Bird	m	PM74	7689

[i]Sample locality (PRA = Provincial Recreational Area, NA = Natural Area, PP = Provincial Park, NF = National Forest, Cpgd = Campground, TWP = township road, RR = range road, N=north, E=east, W=west, S=south).

[ii]Collectors (ASRD = Alberta Sustainable Resource Development crew Erica Lee; AESRD = Alberta Environment and Sustainable Resource Development crew: Oksana Izio, Jade Garland, Jennifer MacCormick, and Dale Thomas at Slake Lake; Kendall Hunt, Mike Maximchuk, Nick Paranko, and Tom Hutchison at Peace River; Devin Letourneau at Grand Prairie; Dave Moseley, and Alex Beauliue at Lac la Biche; Ed Trenchard, and Chris Breen at Whitecourt; and Anina Hundsdörfer at Edmonton; FIDS = Canadian Forest Insect & Disease Survey crew Mike Grandmaison, Robert Sajan, Peter Koot, Janice Hodge, Christian Hébert, Linda Keyes, R. Ferris, O.A.M. (ranger #42), S.R. Cormier, Alan Stewart, and Tom Gray; CFS GLFC = Canadian Forest Service, Great Lakes Forestry Centre)

[iii]Restriction enzyme (A = ApeKI, PM = PstI-MspI) sample number.

[iv]DNA number identifier used in the Sperling Lab.

Table A-2
Number of unique SNPs by species and species combinations.

Clade[i]	ApeKI[ii]	PstI-MspI[ii]
bi	0 (8)	4 (6)
ca	1 (8)	42 (1)
oc	0 (26)[iii]	4 (8)
bi-ca	0 (16)	5 (7)
bi-oc	0 (34)	4 (14)
oc-ca	0 (34)	8 (9)
oc-bi-ca	16 (42)	51 (15)
re	182 (3)	n/a
oc-bi-ca-re	11 (45)	n/a
fu	34 (41)	37 (59)[iv]
re-fu	4 (44)	n/a
oc-bi-ca-fu	24 (83)	56 (74)
oc-bi-ca-re-fu	315 (86)	n/a
pi	594 (8)	353 (65)
oc-bi-ca-pi	3 (50)	103 (80)
re-pi	37 (11)	n/a
oc-bi-ca-re-pi	25 (53)	n/a
fu-pi	10 (49)	17 (124)
oc-bi-ca-fu-pi	122 (91)	117 (138)
re-fu-pi	6 (52)	n/a
oc-bi-ca-re-fu-pi	5214 (94)	n/a
ro	4697 (3)	2245 (3)
co	3830 (2)	104 (2)

[i]Species are identified by the first two letters of their names, oc=*occidentalis*, bi=*biennis*, ca=*carnana*, re=*retiniana*, fu=*fumiferana*, pi=*pinus*, ro=*rosaceana*, and co=*conflictana*.
[ii]Using the read depth threshold 2 and locus filter 75% datasets. The locus filter removes loci based on missing data, according to a minimum proportion of specimens genotyped.
[iii](With the number of specimens in brackets).
[iv]Not including fu x oc hybrid specimen.

Table A-3

Average pair-wise evolutionary divergences within each species, ApeKI (taxa=99, SNPs=789,600) and PstI-MspI analyses (taxa=144, SNPs=201,748).

	ApeKI		PstI-MspI	
	n	**p-distance**[i]	**n**	**p-distance**
C. biennis	8	0.001395	6	0.000822
C. carnana	8	0.001461	1	n/a
C. conflictana	2	0.000125	2	0.000295
C. fumiferana	41	0.001666	59	0.001002
C. occidentalis	26	0.001476	8	0.000797
C. pinus	8	0.001002	65	0.000781
C. retiniana	3	0.000807	0	n/a
C. rosaceana	3	0.000670	3	0.000602

[i] P-distance calculated using MEGA5.1, corrected to include invariant bases, using the read depth threshold 2 and locus filter of a minimum of 10% of specimens genotyped.

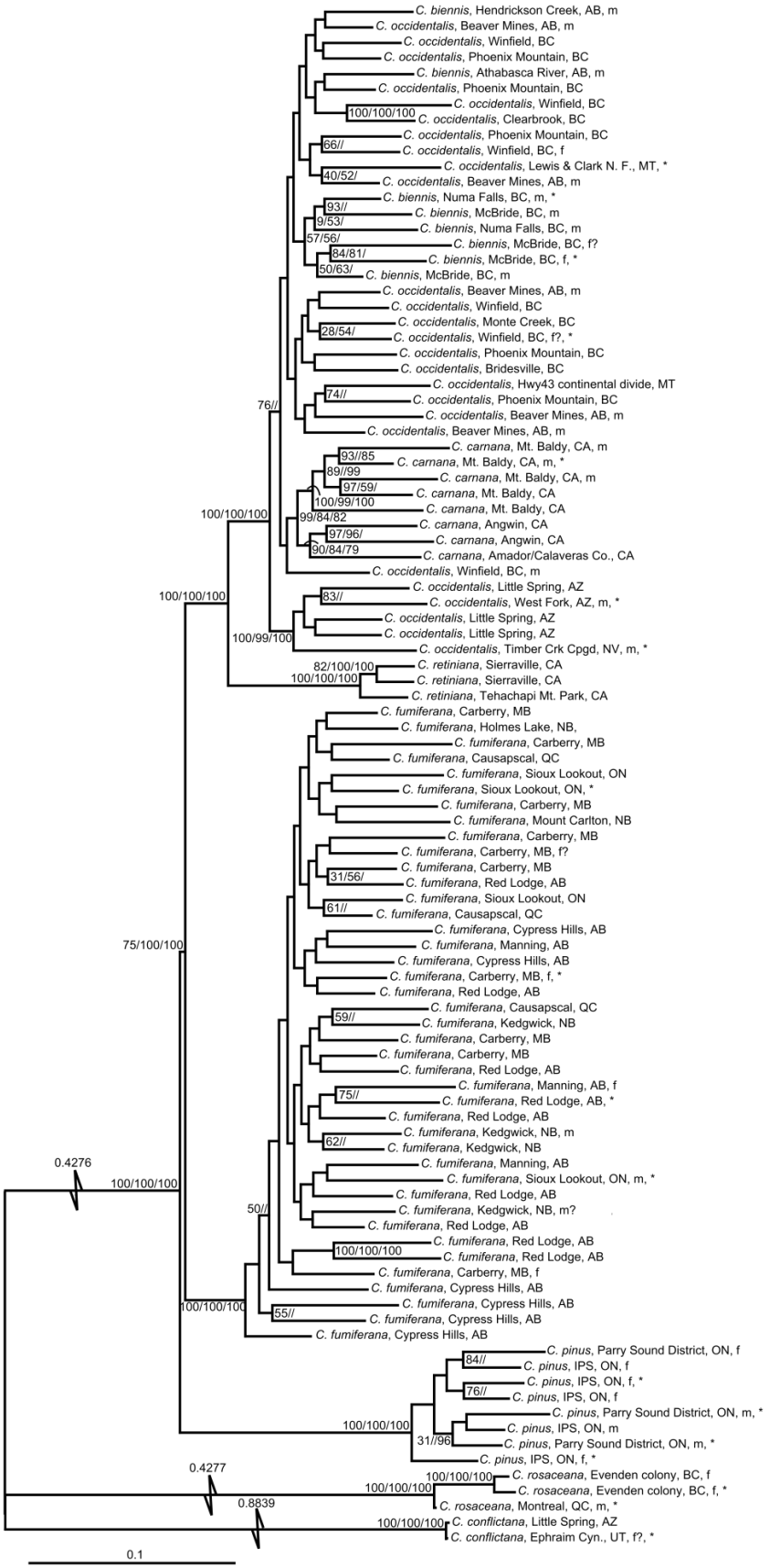
Table A-4
Pair-wise evolutionary divergences between eight *Choristoneura* species.

A\PM[i]	***C. pinus***	***C. fumiferana***	***C. occidentalis***	***C. biennis***	***C. carnana***	***C. retiniana***	***C. rosaceana***	***C. conflictana***
C. pinus		**0.002347**	**0.002251**	**0.002262**	**0.002254**	n/a	**0.003986**	**0.003150**
C. fumiferana	**0.002718**		**0.001565**	**0.001561**	**0.001605**	n/a	**0.003483**	**0.002945**
C. occidentalis	**0.002751**	**0.002296**		0.000836	0.000807	n/a	**0.003486**	**0.002899**
C. biennis	**0.002694**	**0.002199**	0.001459		0.000849	n/a	**0.003568**	**0.002976**
C. carnana	**0.002771**	**0.002317**	**0.001530**	0.001421		n/a	**0.003533**	**0.002952**
C. retiniana	**0.002817**	**0.002498**	**0.001981**	**0.001828**	**0.001968**		n/a	n/a
C. rosaceana	**0.006221**	**0.006219**	**0.006200**	**0.005786**	**0.006231**	**0.006925**		**0.003017**
C. conflictana	**0.007343**	**0.007325**	**0.007344**	**0.006802**	**0.007342**	**0.008229**	**0.006951**	

[i]ApeKI (taxa=99, SNPs=789,600) below diagonal, and PstI-MspI analyses (taxa=144, SNPs=201,748) above diagonal, p-distance calculated using MEGA5.1, corrected to include invariant bases, using the read depth threshold of 2 and locus filter of a minimum of 10% of specimens genotyped. Bold values are significant (p-value < 0.01) calculated using a two-sample t-test comparing pairwise intraspecific distances to interspecific distances of each species and species cross.

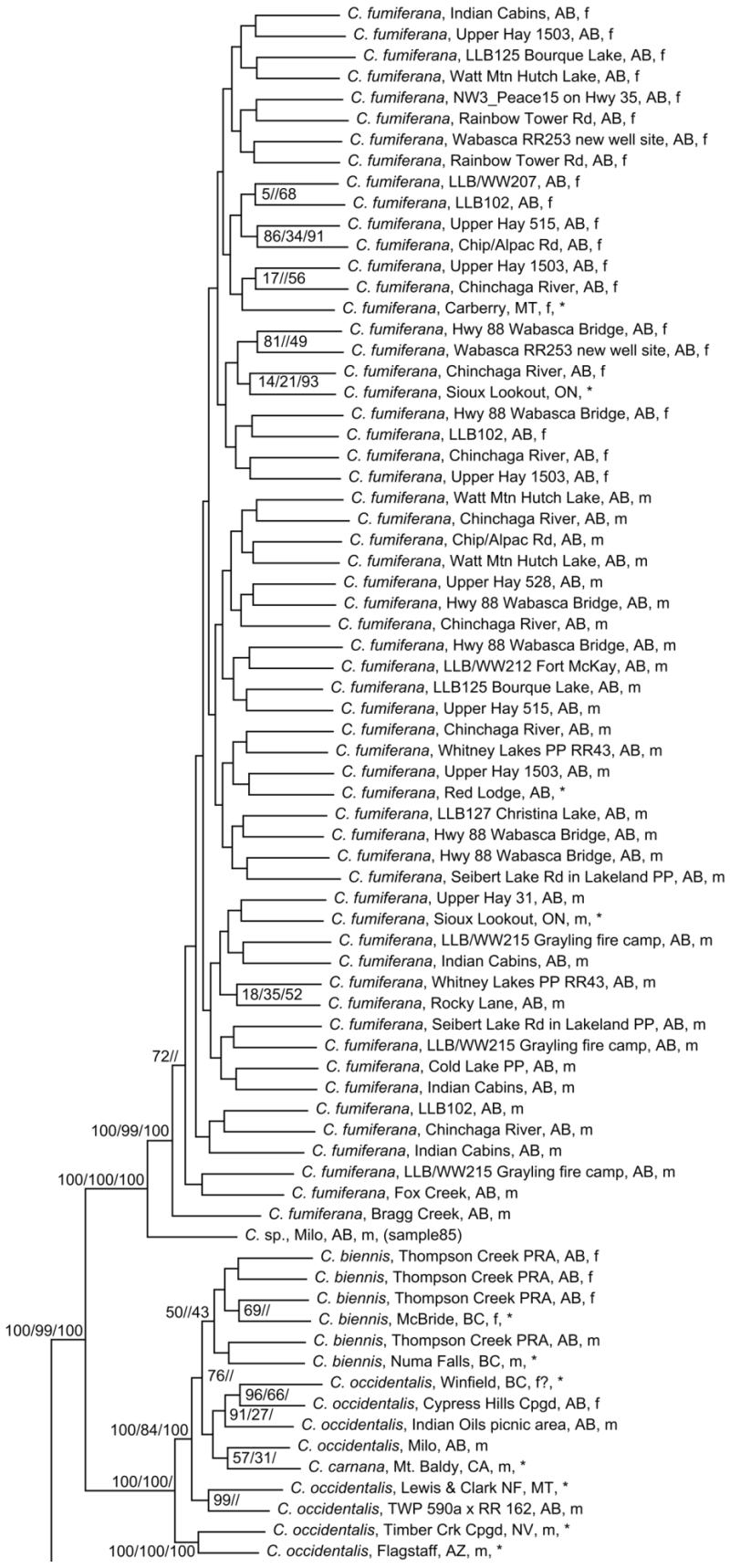
Following page:

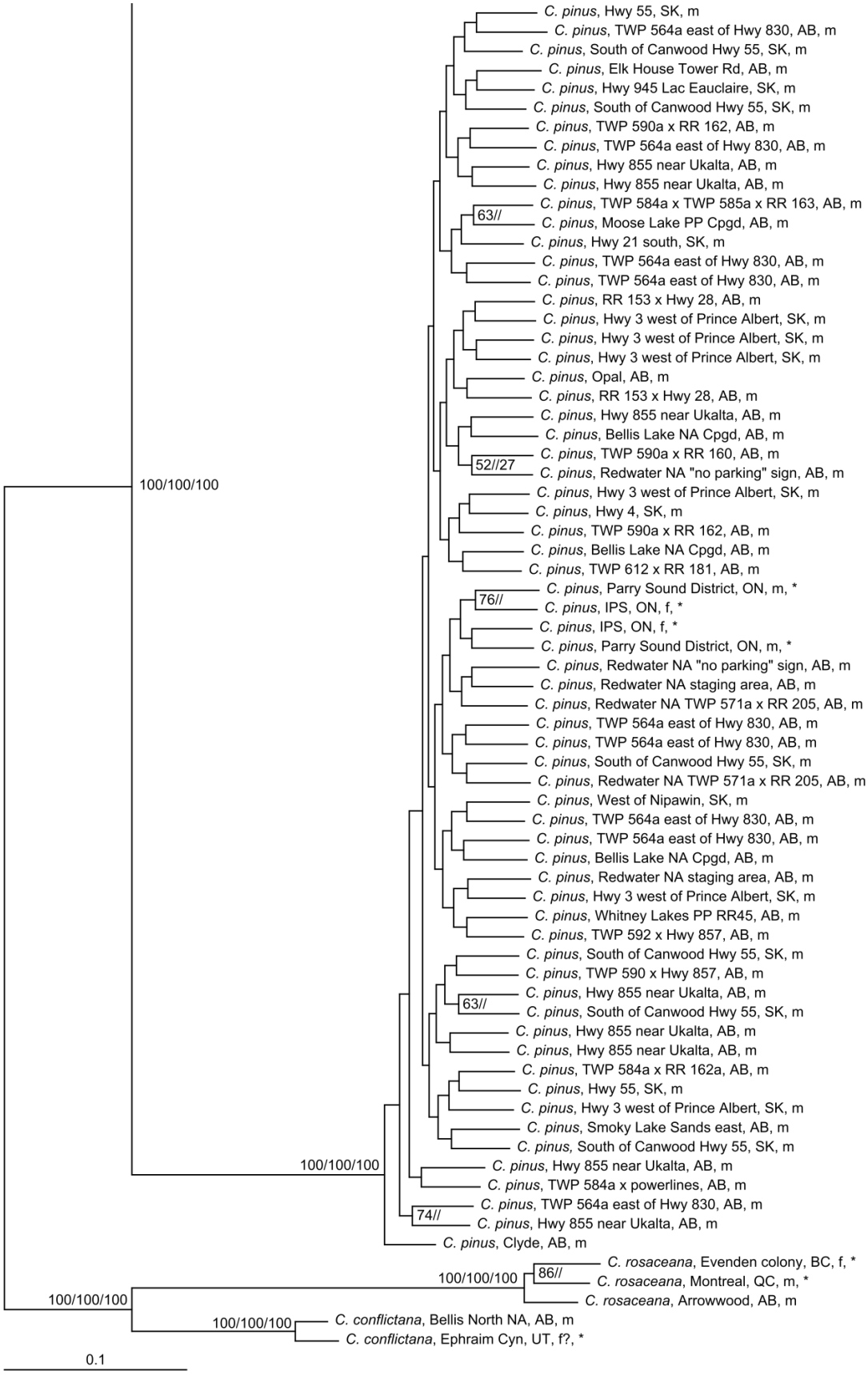
Figure A-1. ApeKI full Maximum Likelihood phylogeny with Maximum Likelihood / Maximum Parsimony / Neighbour Joining bootstrap values. Using 56,760 SNPs with greater than 75% of the 99 specimens genotyped for ML, and the larger dataset of 789,627 SNPs with greater than 10% of specimens genotyped for MP and NJ. Specimens are labeled with species, collection location, sex (m = male, f = female), and a * if that individual was also included in the PstI-MspI analysis.


C. biennis, Hendrickson Creek, AB, m
C. occidentalis, Beaver Mines, AB, m
C. occidentalis, Winfield, BC
C. occidentalis, Phoenix Mountain, BC
C. biennis, Athabasca River, AB, m
C. occidentalis, Phoenix Mountain, BC
C. occidentalis, Winfield, BC
100/100/100
C. occidentalis, Clearbrook, BC
C. occidentalis, Phoenix Mountain, BC
66//
C. occidentalis, Winfield, BC, f
C. occidentalis, Lewis & Clark N. F., MT, *
40/52/
C. occidentalis, Beaver Mines, AB, m
C. biennis, Numa Falls, BC, m, *
93//
C. biennis, McBride, BC, m
9/53/
C. biennis, Numa Falls, BC, m
57/56/
C. biennis, McBride, BC, f?
84/81/
C. biennis, McBride, BC, f, *
50/63/
C. biennis, McBride, BC, m
C. occidentalis, Beaver Mines, AB, m
C. occidentalis, Winfield, BC
C. occidentalis, Monte Creek, BC
28/54/
C. occidentalis, Winfield, BC, f?, *
C. occidentalis, Phoenix Mountain, BC
C. occidentalis, Bridesville, BC
C. occidentalis, Hwy43 continental divide, MT
74//
C. occidentalis, Phoenix Mountain, BC
76//
C. occidentalis, Beaver Mines, AB, m
C. occidentalis, Beaver Mines, AB, m
C. carnana, Mt. Baldy, CA, m
93//85
C. carnana, Mt. Baldy, CA, m, *
89//99
C. carnana, Mt. Baldy, CA, m
97/59/
C. carnana, Mt. Baldy, CA
100/99/100
C. carnana, Mt. Baldy, CA
100/100/100
99/84/82
C. carnana, Angwin, CA
97/96/
C. carnana, Angwin, CA
90/84/79
C. carnana, Amador/Calaveras Co., CA
C. occidentalis, Winfield, BC, m
C. occidentalis, Little Spring, AZ
100/100/100
83//
C. occidentalis, West Fork, AZ, m, *
C. occidentalis, Little Spring, AZ
100/99/100
C. occidentalis, Little Spring, AZ
C. occidentalis, Timber Crk Cpgd, NV, m, *
82/100/100
C. retiniana, Sierraville, CA
100/100/100
C. retiniana, Sierraville, CA
C. retiniana, Tehachapi Mt. Park, CA
C. fumiferana, Carberry, MB
C. fumiferana, Holmes Lake, NB,
C. fumiferana, Carberry, MB
C. fumiferana, Causapscal, QC
C. fumiferana, Sioux Lookout, ON
C. fumiferana, Sioux Lookout, ON, *
C. fumiferana, Carberry, MB
C. fumiferana, Mount Carlton, NB
C. fumiferana, Carberry, MB
C. fumiferana, Carberry, MB, f?
C. fumiferana, Carberry, MB
31/56/
C. fumiferana, Red Lodge, AB
C. fumiferana, Sioux Lookout, ON
61//
C. fumiferana, Causapscal, QC
C. fumiferana, Cypress Hills, AB
75/100/100
C. fumiferana, Manning, AB
C. fumiferana, Cypress Hills, AB
C. fumiferana, Carberry, MB, f, *
C. fumiferana, Red Lodge, AB
C. fumiferana, Causapscal, QC
59//
C. fumiferana, Kedgwick, NB
C. fumiferana, Carberry, MB
C. fumiferana, Carberry, MB
C. fumiferana, Red Lodge, AB
C. fumiferana, Manning, AB, f
75//
C. fumiferana, Red Lodge, AB, *
C. fumiferana, Red Lodge, AB
C. fumiferana, Kedgwick, NB, m
62//
C. fumiferana, Kedgwick, NB
0.4276
C. fumiferana, Manning, AB
100/100/100
C. fumiferana, Sioux Lookout, ON, m, *
C. fumiferana, Red Lodge, AB
50//
C. fumiferana, Kedgwick, NB, m?
C. fumiferana, Red Lodge, AB
C. fumiferana, Red Lodge, AB
100/100/100
C. fumiferana, Red Lodge, AB
C. fumiferana, Carberry, MB, f
C. fumiferana, Cypress Hills, AB
100/100/100
C. fumiferana, Cypress Hills, AB
55//
C. fumiferana, Cypress Hills, AB
C. fumiferana, Cypress Hills, AB
C. pinus, Parry Sound District, ON, f
84//
C. pinus, IPS, ON, f
C. pinus, IPS, ON, f, *
76//
C. pinus, IPS, ON, f
C. pinus, Parry Sound District, ON, m, *
100/100/100
C. pinus, IPS, ON, m
31//96
C. pinus, Parry Sound District, ON, m, *
C. pinus, IPS, ON, f, *
0.4277
100/100/100
C. rosaceana, Evenden colony, BC, f
100/100/100
C. rosaceana, Evenden colony, BC, f, *
0.8839
C. rosaceana, Montreal, QC, m, *
100/100/100
C. conflictana, Little Spring, AZ
C. conflictana, Ephraim Cyn., UT, f?, *
0.1

Following two pages:

Figure A-2. PstI-MspI full Maximum Likelihood phylogeny with Maximum Likelihood / Maximum Parsimony / Neighbour bootstrap values. Using 57,440 SNPs with greater than 75% of the 144 specimens genotyped for ML, and the larger dataset of 201,791 SNPs with greater than 10% of specimens genotyped for MP and NJ. Specimens are labeled with species, collection location, sex (m = male, f = female), and a * if that individual was also included in the ApeKI analysis.


C. fumiferana, Indian Cabins, AB, f
C. fumiferana, Upper Hay 1503, AB, f
C. fumiferana, LLB125 Bourque Lake, AB, f
C. fumiferana, Watt Mtn Hutch Lake, AB, f
C. fumiferana, NW3_Peace15 on Hwy 35, AB, f
C. fumiferana, Rainbow Tower Rd, AB, f
C. fumiferana, Wabasca RR253 new well site, AB, f
C. fumiferana, Rainbow Tower Rd, AB, f
C. fumiferana, LLB/WW207, AB, f
5//68
C. fumiferana, LLB102, AB, f
C. fumiferana, Upper Hay 515, AB, f
86/34/91
C. fumiferana, Chip/Alpac Rd, AB, f
C. fumiferana, Upper Hay 1503, AB, f
17//56
C. fumiferana, Chinchaga River, AB, f
C. fumiferana, Carberry, MT, f, *
C. fumiferana, Hwy 88 Wabasca Bridge, AB, f
81//49
C. fumiferana, Wabasca RR253 new well site, AB, f
C. fumiferana, Chinchaga River, AB, f
14/21/93
C. fumiferana, Sioux Lookout, ON, *
C. fumiferana, Hwy 88 Wabasca Bridge, AB, f
C. fumiferana, LLB102, AB, f
C. fumiferana, Chinchaga River, AB, f
C. fumiferana, Upper Hay 1503, AB, f
C. fumiferana, Watt Mtn Hutch Lake, AB, m
C. fumiferana, Chinchaga River, AB, m
C. fumiferana, Chip/Alpac Rd, AB, m
C. fumiferana, Watt Mtn Hutch Lake, AB, m
C. fumiferana, Upper Hay 528, AB, m
C. fumiferana, Hwy 88 Wabasca Bridge, AB, m
C. fumiferana, Chinchaga River, AB, m
C. fumiferana, Hwy 88 Wabasca Bridge, AB, m
C. fumiferana, LLB/WW212 Fort McKay, AB, m
C. fumiferana, LLB125 Bourque Lake, AB, m
C. fumiferana, Upper Hay 515, AB, m
C. fumiferana, Chinchaga River, AB, m
C. fumiferana, Whitney Lakes PP RR43, AB, m
C. fumiferana, Upper Hay 1503, AB, m
C. fumiferana, Red Lodge, AB, *
C. fumiferana, LLB127 Christina Lake, AB, m
C. fumiferana, Hwy 88 Wabasca Bridge, AB, m
C. fumiferana, Hwy 88 Wabasca Bridge, AB, m
C. fumiferana, Seibert Lake Rd in Lakeland PP, AB, m
C. fumiferana, Upper Hay 31, AB, m
C. fumiferana, Sioux Lookout, ON, m, *
C. fumiferana, LLB/WW215 Grayling fire camp, AB, m
C. fumiferana, Indian Cabins, AB, m
C. fumiferana, Whitney Lakes PP RR43, AB, m
18/35/52
C. fumiferana, Rocky Lane, AB, m
C. fumiferana, Seibert Lake Rd in Lakeland PP, AB, m
72//
C. fumiferana, LLB/WW215 Grayling fire camp, AB, m
C. fumiferana, Cold Lake PP, AB, m
C. fumiferana, Indian Cabins, AB, m
C. fumiferana, LLB102, AB, m
100/99/100
C. fumiferana, Chinchaga River, AB, m
C. fumiferana, Indian Cabins, AB, m
100/100/100
C. fumiferana, LLB/WW215 Grayling fire camp, AB, m
C. fumiferana, Fox Creek, AB, m
C. fumiferana, Bragg Creek, AB, m
C. sp., Milo, AB, m, (sample85)
C. biennis, Thompson Creek PRA, AB, f
C. biennis, Thompson Creek PRA, AB, f
50//43
C. biennis, Thompson Creek PRA, AB, f
69//
C. biennis, McBride, BC, f, *
100/99/100
C. biennis, Thompson Creek PRA, AB, m
C. biennis, Numa Falls, BC, m, *
76//
C. occidentalis, Winfield, BC, f?, *
96/66/
C. occidentalis, Cypress Hills Cpgd, AB, f
91/27/
C. occidentalis, Indian Oils picnic area, AB, m
100/84/100
C. occidentalis, Milo, AB, m
57/31/
C. carnana, Mt. Baldy, CA, m, *
100/100/
C. occidentalis, Lewis & Clark NF, MT, *
99//
C. occidentalis, TWP 590a x RR 162, AB, m
C. occidentalis, Timber Crk Cpgd, NV, m, *
100/100/100
C. occidentalis, Flagstaff, AZ, m, *


C. pinus, Hwy 55, SK, m
C. pinus, TWP 564a east of Hwy 830, AB, m
C. pinus, South of Canwood Hwy 55, SK, m
C. pinus, Elk House Tower Rd, AB, m
C. pinus, Hwy 945 Lac Eauclaire, SK, m
C. pinus, South of Canwood Hwy 55, SK, m
C. pinus, TWP 590a x RR 162, AB, m
C. pinus, TWP 564a east of Hwy 830, AB, m
C. pinus, Hwy 855 near Ukalta, AB, m
C. pinus, Hwy 855 near Ukalta, AB, m
C. pinus, TWP 584a x TWP 585a x RR 163, AB, m
63//
C. pinus, Moose Lake PP Cpgd, AB, m
C. pinus, Hwy 21 south, SK, m
C. pinus, TWP 564a east of Hwy 830, AB, m
C. pinus, TWP 564a east of Hwy 830, AB, m
C. pinus, RR 153 x Hwy 28, AB, m
C. pinus, Hwy 3 west of Prince Albert, SK, m
C. pinus, Hwy 3 west of Prince Albert, SK, m
C. pinus, Hwy 3 west of Prince Albert, SK, m
C. pinus, Opal, AB, m
C. pinus, RR 153 x Hwy 28, AB, m
C. pinus, Hwy 855 near Ukalta, AB, m
C. pinus, Bellis Lake NA Cpgd, AB, m
C. pinus, TWP 590a x RR 160, AB, m
52//27
C. pinus, Redwater NA "no parking" sign, AB, m
100/100/100
C. pinus, Hwy 3 west of Prince Albert, SK, m
C. pinus, Hwy 4, SK, m
C. pinus, TWP 590a x RR 162, AB, m
C. pinus, Bellis Lake NA Cpgd, AB, m
C. pinus, TWP 612 x RR 181, AB, m
C. pinus, Parry Sound District, ON, m, *
76//
C. pinus, IPS, ON, f, *
C. pinus, IPS, ON, f, *
C. pinus, Parry Sound District, ON, m, *
C. pinus, Redwater NA "no parking" sign, AB, m
C. pinus, Redwater NA staging area, AB, m
C. pinus, Redwater NA TWP 571a x RR 205, AB, m
C. pinus, TWP 564a east of Hwy 830, AB, m
C. pinus, TWP 564a east of Hwy 830, AB, m
C. pinus, South of Canwood Hwy 55, SK, m
C. pinus, Redwater NA TWP 571a x RR 205, AB, m
C. pinus, West of Nipawin, SK, m
C. pinus, TWP 564a east of Hwy 830, AB, m
C. pinus, TWP 564a east of Hwy 830, AB, m
C. pinus, Bellis Lake NA Cpgd, AB, m
C. pinus, Redwater NA staging area, AB, m
C. pinus, Hwy 3 west of Prince Albert, SK, m
C. pinus, Whitney Lakes PP RR45, AB, m
C. pinus, TWP 592 x Hwy 857, AB, m
C. pinus, South of Canwood Hwy 55, SK, m
C. pinus, TWP 590 x Hwy 857, AB, m
C. pinus, Hwy 855 near Ukalta, AB, m
63//
C. pinus, South of Canwood Hwy 55, SK, m
C. pinus, Hwy 855 near Ukalta, AB, m
C. pinus, Hwy 855 near Ukalta, AB, m
C. pinus, TWP 584a x RR 162a, AB, m
C. pinus, Hwy 55, SK, m
C. pinus, Hwy 3 west of Prince Albert, SK, m
C. pinus, Smoky Lake Sands east, AB, m
C. pinus, South of Canwood Hwy 55, SK, m
100/100/100
C. pinus, Hwy 855 near Ukalta, AB, m
C. pinus, TWP 584a x powerlines, AB, m
C. pinus, TWP 564a east of Hwy 830, AB, m
74//
C. pinus, Hwy 855 near Ukalta, AB, m
C. pinus, Clyde, AB, m
C. rosaceana, Evenden colony, BC, f, *
100/100/100
86//
C. rosaceana, Montreal, QC, m, *
C. rosaceana, Arrowwood, AB, m
100/100/100
100/100/100
C. conflictana, Bellis North NA, AB, m
C. conflictana, Ephraim Cyn, UT, f?, *
0.1

2.5. References

Abe, H., Mita, K., Yasukochi, Y., Oshiki, T., Shimada, T. 2005. Retrotransposable elements on the W chromosome of the silkworm, *Bombyx mori*. Cytogenet. Genome Res. 110, 144-151.

Anderson, L.L., Hu, F.S., Paige, K.N. 2011. Phylogeographic history of white spruce during the last glacial maximum: uncovering cryptic refugia. J. Hered. 102, 207-216.

Andrew, R.L., Bernatchez, L., Bonin, A., Buerkle, C.A., Carstens, B.C., Emerson, B.C., Garant, D., Giraud, T., Kane, N.C., Rogers, S.M., Slate, J., Smith, H., Sork, V.L., Stone, G.N., Vines, T.H., Waits, L., Widmer, A. Rieseberg, L.H. 2013. A road map for molecular ecology. Mol. Ecol. 22, 2605-2626.

Aquadro, C.F., DuMont, V.B., Reed, F.A. 2001 Genome-wide variation in the human and fruitfly: a comparison. Curr. Opin. Genet. Dev. 11, 627-634.

Arnold, B., Corbett-Detif, R.B., Hartl, D., Bomblies, K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. 22, 3179-3190.

Beacham, T.D., Wetklo, M., Wallace, C., Olsen, J.B., Flannery, B.G., Wenburg, J.K., Templin, W.D., Antonovich, A., Seeb, L.W. 2008. The application of microsatellites for stock identification of Yukon River Chinook salmon. N. Am. J. Fish. Manage. 28, 283-295.

Bossu, C.M., Near, T.J. 2009. Gene trees reveal repeated instances of mitochondrial DNA introgression in orangethroat darters (Percidae: Etheostoma). Syst. Biol. 58, 114-129.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 23, 2633-2635.

Brown, J.W., Baixeras, J., Brown, R., Horak, M., Komai, F., Metzler, E., Razowski, J., Tuck, K. 2005. World Catalogue of Insects, Volume 5, Tortricidae (Lepidoptera). Apollo Books, Stenstrup, Denmark.

Brown, B., Emberson, R.M., Paterson, A.M. 1999. Phylogeny of "*Oxycanus*" lineages of Hepialid moths from New Zealand inferred from sequence variation in the mtDNA COI and II gene regions. Mol. Phylogenet. Evol. 13, 463-473.

Campbell, I.M. 1967. On coniferophagous species of *Choristoneura* (Lepidoptera: Tortricidae) in North America. IV. Sexual Isolation between three species. Can. Entomol. 99, 482-486.

Cariou, M., Duret, L., Charlat, S. 2013. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. Ecol. Evol. 3, 846-852.

Castel, A.L., Nakamori, M., Thornton, C.A., Pearson, C.E. 2011. Identification of restriction endonucleases sensitive to 5-cytosine methylation at non-CpG sites, including expanded (CAG)n/(CTG)n repeats. Epigenetics. 6, 416-420.

Castrovillo, P.J. 1982. Interspecific and intraspecific genetic comparisons of North American spruce budworms (*Choristoneura* spp.). PhD dissertation. University of Idaho.

Colbourne, J.K., Pfrender, M.E., Gilbert, et al., (66 co-authors). 2011. The ecoresponsive genome of *Daphnia pulex*. Science. 331, 555–561.

Corl, A., Ellegren, H. 2013. Sampling strategies for species trees: The effects on phylogenetic inference of the number of genes, number of individuals, and whether loci are mitochondrial, sex-linked, or autosomal. Mol. Phylogenet. Evol. 67, 358-366.

Davey, J.W., Blaxter, M.L. 2010. RADSeq: next-generation population genetics. Brief. Funct. Genomics. 9, 416-423.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12, 499-510.

Dombroskie, J.J. 2011. Aspects of archipine evolution (Lepidoptera: Tortricidae). Ph.D. thesis. University of Alberta. Department of Biological Sciences. Edmonton, Alberta, Canada.

Dupuis, J.R., Roe, A.D., Sperling, F.A.H. 2012. Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. Mol. Ecol. 21, 4422-4436.

Eaton, D.A.R., Ree, R.H. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol. 62, 689-706.

Edwards, S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1-19.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 6, e19379.

Emerson K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E., Holzapfel, C.M. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl. Acad. Sci. 107, 16196-16200.

Ersts, P.J. 2013. Geographic Distance Matrix Generator (version 1.2.3). American Museum of Natural History, Center for Biodiversity and Conservation. Available from http://biodiversityinformatics.amnh.org/open_source/gdmg. Accessed on 2013-3-25.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 39, 783-791.

Fitzpatrick, B.M., Fordyce, J.A., Gavrilets, S. 2008. What, if anything, is sympatric speciation? J. Evol. Biol. 21, 1452-1459.

Freeman, T.N. 1953. The spruce budworm, *Choristoneura fumiferana* (Clem.) and an allied new species on pine (Lepidoptera: Tortricidae). Can. Entomol. 85, 121-127.

Freeman, T.N. 1967. On coniferophagous species of *Choristoneura* (Lepidoptera: Tortricidae) in North America I. Some new forms of *Choristoneura* allied to *C. fumiferana*. Can. Entomol. 99, 449-455.

Fuková, I., Nguyen, P., Marec, F. 2005. Codling moth cytogenetics: karyotype, chromosomal location of rDNA, and molecular differentiation of sex chromosomes. Genome. 48. 1083-1092.

Funk, D.J., Omland, K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34, 397-423.

Futuyma, D.J., Peterson, S.C. 1985. Genetic variation in the use of resources by insects. Ann. Rev. Entomol. 30, 217-238.

Gatesy, J., DeSalle, R., Wahlberg, N. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst. Biol. 56. 355-363.

Glaubitz, J., Harriman, J., Casstevens, T. 2012. TASSEL 3.0 Genotyping By Sequencing (GBS) pipeline documentation. Available from www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf. Accessed 2012-2-24.

Glenn, T.C. 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 11, 759-769.

Gompert, Z., Forister, M.L., Fordyce, J.A., Nice, C.C., Williamson, R.J., Buerkle, C.A. 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. Mol. Ecol. 19, 2455-2473.

Gong, Y.J., Shi, B.C., Kang, Z.J., Zhang, F. Wei, S.J. 2012. The complete mitochondrial genome of the oriental fruit moth Grapholita molesta (Busck) (Lepidoptera: Tortricidae). Mol. Biol. Rep. 36, 2893-2900.

Gruenbaum, Y., Cedar, H., Razin, A. 1981. Restriction enzyme digestion of hemimethylated DNA. Nucleic Acids Res. 9, 2509-2515.

Gugger, P.F., Sugita, S. 2010. Glacial populations and postglacial migration of Douglas-fir based on fossil pollen and macrofossil evidence. Quat. Sci. Rev. 29, 2052-2070.

Hare, E.E, Johnston, J.S. 2011. Genome Size Determination using Flow Cytometry of Propidium Iodide-stained Nuclei. In: Orgogozo, V.,

Rockman, M.V. (Eds.), Methods Molecular Biology, Volume 772. Humana Press, Springer Science and Business Media, LLC.

Harvey, G.T. 1967. On coniferophagous species on *Choristoneura* (Lepidoptera: Tortricidae) in North America. V. Second diapause as a species character. Can. Entomol. 99, 486-503.

Harvey, G.T. 1983. Geographical cline in egg weights in *Choristoneura fumiferana* (Lepidoptera: Tortricidae) and its significance in population dynamics. Can. Entomol. 115, 1103–1108.

Harvey, G.T. 1985. The taxonomy of the Coniferophagous *Choristoneura* (Lepidoptera: Tortricidae): A review. In: Saunders, C.L., Stark, R.W., Mullins, E.J., Murphy, J. (Eds.), Recent Advances in Spruce Budworm Research. Proceedings CANUSA Spruce budworms Research Symposium, Bangor, ME. Canadian Forest Service, Ottawa, Canada, pp. 16-48.

Harvey, G.T. 1996. Genetic relationships among *Choristoneura* species (Lepidoptera: Tortricidae) in North America as revealed by isozyme studies. Can. Entomol. 128, 245-262.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., Cresko, W.A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6, e1000862.

Hohenlohe, P.A., Amish, S. J., Catchen, J.M., Allendorf, F.W., Luikart, G. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol. Ecol. Resour. 11, 117-122.

Hohenlohe, P.A., Catchen, J., Cresko, W.A. 2012. Population Genomic Analysis of Model and Nonmodel Organisms Using Sequenced RAD Tags. In: Pompanon, F., Bonin, A. (Eds.), Data Production and Analysis in Population Genomics: Methods and Protocols. Humana Press, Springer Science, New York.

Jaenisch, R., Bird, A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat. Genet. 33, 245-254.

Jones, J.C., Fan, S., Franchini, P., Schartl, M., Meyer, A. 2013. The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. Mol. Ecol. early view, mec.12269.

Karlin, S., Ladunga, I. 1994. Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. USA. 91, 12832-12836.

Kawaoka, S., Kadota, K., Arai, Y., Suzuki, Y., Fujii, T., Abe, H., Yasukochi, Y., Mita, K., Sugano, S., Shimizu, K., Tomari, Y., Shimada, T., Katsuma, S. 2011. The silkworm W chromosome, is a source of female-enriched piRNAs. RNA. 17, 2144-2151.

Lee, E.S., Shin, K.S., Kim, M.S., Park, H., Cho, S., Kim, C.B. 2006. The mitochondrial genome of the smaller tea tortrix *Adoxophyes honmai* (Lepidoptera: Tortricidae). Gene. 373, 52-57.

Li, H., Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 25, 1754-60.

Lindroth, R.L. 1991. Biochemical ecology of Aspen-Lepidoptera interactions. J. Kans. Entomol. Soc. 64, 372-380.

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S., Costich, D.E. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genet. 9, e1003215.

Lumley, L.M. Sperling, F.A.H. 2010. Integrating morphology and mitochondrial DNA for species delimitation within the spruce budworm (*Choristoneura fumiferana*) cryptic species complex (Lepidoptera: Tortricidae). Syst. Entomol. 35, 416-428.

Lumley, L.M., Sperling, F.A.H. 2011a. Utility of microsatellites and mitochondrial DNA for species delimitation in the spruce budworm

(*Choristoneura fumiferana*) species complex (Lepidoptera: Tortricidae). Mol. Phylogenet. Evol. 58, 232-243.

Lumley, L.M., Sperling, F.A.H. 2011b. Life-history traits maintain the genomic integrity of sympatric species of the spruce budworm (*Choristoneura fumiferana*) on an isolated forest island. Ecol. Evol. 1, 119-133.

Maddison, D.R., Ober, K.A. 2011. Phylogeny of minute carabid beetles and their relatives based upon DNA sequence data (Coleoptera, Carabidae, Trechitae). Zookeys. 147, 229-260.

Maddison, W.P., Knowles, L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21-30.

Maine Forest Service. 2000. Spruce budworm *Choristoneura fumiferana* (Clem.) on ornamentals and Christmas trees. Maine Department of Conversation factsheets, http://www.maine.gov/doc/mfs/documents/sprucebudworm.pdf, Retrieved April 24, 2013.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20, 229-237.

McCormack, J.E., Maley, J.M., Hird, S.M., Derryberry, E.P., Graves, G.R., Brumfield, R.T. 2012. Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. Mol. Phylogenet. Evol. 62, 397-406.

Metzker, M.L. 2010. Sequencing technologies – the next generation. Nat. Rev. Genet. 11, 31-46.

Meusemann, K., von Reumont, B.M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., von Haeseler, A., Burmester, T., Hadrys, H., Wägele, J.,W., Misof, B. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27, 2451-2464.

Nealis, V.G. 2005. Diapause and voltinism in western and 2-year-cycle spruce budworms (Lepidoptera: Tortricidae) and their hybrid progeny. Can. Entomol. 137, 584-597.

Nei, M., Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nice, C.C., Fordyce, J.A., Shapiro, A.M., Ffrench-Constant, R. 2002. Lack of evidence for reproductive isolation among ecologically specialised lycaenid butterflies. Ecol. Entomol. 27, 702-712.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., Wang, J. 2012. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS ONE. 7, e37558.

Nguyen, P., Sykorová, M., Síchová, J., Kuta, V., Dalíková, M., Capková Frydrychová, R., Neven, L.G., Sahara, K., Marec, F. 2013. Neo-sex chromosomes and adaptive potential in tortricid pests. Proc. Natl. Acad. Sci. 110, 6931-6936.

Pashley, D.P., Ke, L.D. 1992. Sequence evolution in mitochondrial ribosomal and ND-1 genes in Lepidoptera: implications for phylogenetic analyses. Mol. Biol. Evol. 9, 1061-1075.

Pielou, E.C. 1991. After the Ice Age: The Return of Life to Glaciated North America. University of Chicago Press, Chicago, IL, USA.

Poland, J.A., Brown, P.J., Sorrells, M.E., Jannink, J.-L. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7, e32253.

Powell, J.A., De Benedictus, J.A. 1995. Evolutionary interpretation, taxonomy and nomenclature. In: Powell, J.A. (Ed.), Biosystematic Studies of Conifer-feeding *Choristoneura* (Lepidoptera: Tortricidae) in the Western United States. University of California Press, Berkeley, CA, USA, pp.219–275.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 13, 341.

Regier, J.C., Brown, J.W., Mitter, C., Baixeras, J., Cho, S., Cummings, M.P., Zwick, A. 2012. A molecular phylogeny for the leaf-roller moths

(Lepidoptera: Tortricidae) and its implications for classification and life history evolution. PLoS ONE. 7, e35574.

Roe, A.D., Sperling, F.A.H. 2007. Patterns of evolution and mitochondrial cytochrome *c* oxidase I and II DNA and the implications for DNA barcoding. Mol. Phylogenet. Evol. 44, 325-345.

Ronquist, F., Huelsenbeck, J., Teslenko, M. 2011. Draft MrBayes version 3.2 Manual: Tutorials and Model Summaries. Available from http://mrbayes.sourceforge.net/mb3.2_manual.pdf . Accessed 2013-09-17.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539-542.

Rüber, L., Meyer, A., Sturmbauer, C., Verheyen, E. 2001. Population structure in two sympatric species of the Lake Tanganyika cichlid tribe Eretmodini: evidence for introgression. Mol. Ecol. 10, 1207-1225.

Rudolph, T.D., Yeatman, C.W. 1982. Genetics of jack pine. USDA Forest Service, Research Paper WO-38. Washington, DC. 60 p.

Sadri, R., Hornsby, P.J. 1996. Rapid analysis of DNA methylation using new restriction enzyme sites created by bisulfite modification. Nucl. Acids Res. 24:5058-5059.

Sanders, C.J. 1974. Sex pheromone specificity and taxonomy of budworm moths (*Choristoneura*). In: Struble, D.L., Brady, U.E. (Eds.), Pheromones: Current Research, Volume 1. Ardent Media.

Sanders, C.J., Daterman, G.E., Ennis, T.J. 1977. Sex pheromone responses of *Choristoneura* spp. and their hybrids (Lepidoptera: Tortricidae). Can. Entomol. 109, 1203-1220.

Sezonlin, M., Dupas, S., Le Rü, B., Le Gall, P., Moyal, P., Calatayud, P.-A., Giffard, I., Faure, N., Silvain, J.-F. 2006. Phylogeography and population genetics of the maize stalk borer *Busseola fusca* (Lepidoptera, Noctuidae) in sub-Saharan Africa. Mol. Ecol. 15, 407-420.

Staubach, F., Lorenc, A., Messer, P.W., Tang, K., Petrov, D.A., Tautz, D. 2012. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). PLoS Genet. 8, e1002891.

Stock, M.W., Castrovillo, P.J. 1981. Genetic relationships among representative populations of five *Choristoneura* species: *C. occidentalis*, *C. retiniana*, *C. biennis*, *C. lambertiana*, and *C. fumiferana* (Lepidoptera: Tortricidae). Can. Entomol. 113, 857-865.

Silk, P.J., Kuenen, L.P.S. 1988. Sex pheromones and behavioral biology of the coniferophagus *Choristoneura*. Ann. Rev. Entomol. 33, 83-101.

Silvestro, D., Michalak, I. 2012. raxmlGUI: A graphical front-end for RAxML. Org. Divers. Evol. 12, 335-337.

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., Normandaeu, É., Laroche, J., Larose, S., Jean, M., Belzile, F. 2013. An improved Genotyping by Sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE. 8, e54603.

Sperling, F.A.H. 1994. Sex-linked genes and species differences in Lepidoptera. Can. Entomol. 126, 807-818.

Sperling, F.A.H., Hickey, D.A. 1994. Mitochondrial DNA sequence variation in the spruce budworm species complex (*Choristoneura*: Lepidoptera) Mol. Biol. Evol. 11, 656-665.

Sperling, F.A.H., Hickey, D.A. 1995. Amplified mitochondrial DNA as a diagnostic marker for species of conifer-feeding *Choristoneura* (Lepidoptera: Tortricidae). Can. Entomol. 127, 277-288.

Stamatakis, A. 2006a. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22, btl446.

Stamatakis, A. 2006b. Phylogenetic models of rate heterogeneity: A high performance computing perspective. In: Proc. of IPDPS2006, Rhodos, Greece.

Stamatakis, A. 2008. The RAxML 7.0.4 Manual. Available from sco.h-its.org/exelixis/oldPage /RAxML-Manual.7.0.4.pdf. Accessed 2013-09-17.

Sturtevant, B.R., Achtemeier, G.L., Charney, J.J., Anderson, D.P., Cooke, B.J., Townsend, P.A. 2013. Long-distance dispersal of spruce budworm (*Choristoneura fumiferana* Clemens) in Minnesota (USA) and Ontario (Canada) via the atmospheric pathway. Agr. Forest Meteorol. 168, 186-200.

Taft, R.J., Mattick, J.S. 2003. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. Genome Biology Preprint Depository, available from http://genomebiology.com/2003/5/1/P1. Accessed 2013-09-17.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28: 2731-2739.

Triplett, J.K., Oltrogge, K.A., Clark, L.G. 2010. Phylogenetic relationships and natural hybridization among the North American woody bamboos (Poaceae: Bambusoideae: *Arundinaria*). Am. J. Bot. 97, 471-492.

Vakenti, J.M., Cossentine, J.E., Cooper, B.E., Sharkey, M.J., Yoshimoto, C.M., Jensen, L.B.M. 2001. Host-plant range and parasitoids of obliquebanded and three-lined leafrollers (Lepidoptera: Tortricidae). Can. Entomol. 133, 139-146.

Volney, W.J.A. 1985. Comparative population biologies of North American spruce budworms. In: Saunders, C.L., Stark, R.W., Mullins, E.J., Murphy, J. (Eds.), Recent Advances in Spruce Budworm Research. Proceedings CANUSA Spruce budworms Research Symposium, Bangor, ME. Canadian Forest Service, Ottawa, Canada, pp. 71-84.

Volney, W.J.A., and Fleming, R.A. 2000. Climate change and impacts of boreal forest insects. Agric. Ecosyst. Environ. 82, 283-294.

Volney, W.J.A., Fleming, R.A. 2007. Spruce budworm (*Choristoneura* spp.) biotype reactions to forest and climate characteristics. Glob. Chang. Biol. 13, 1630-1643.

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., Sivasunder, A., Seehausen, O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787-798.

Walsh, A.M., Kortschak, R.D., Gardner, M.G., Bertozzi, T., Adelson, D.L. 2012. Widespread horizontal transfer of retrotransposons. Proc. Natl. Acad. Sci. 110, 1012-1016.

Wang, X.-P., Yang, G.-J. 2008. A new species of *Choristoneura* Lederer, with a key to the species from China (Lepidoptera: Tortricidae: Tortricinae). Zootaxa. 1944, 66-68.

Weber, U.M., Schweingruber, F.H. 1995. A dendroecological reconstruction of western spruce budworm outbreaks (*Choristoneura occidentalis*) in the Front Range, Colorado, from 1720 to 1986. Trees. 9, 204-213.

Whitehouse, C.M., Roe, A.D., Strong, W.B., Evenden, M.L., Sperling, F.A.H. 2011. Biology and management of North American cone-feeding *Dioryctria* species. Can. Entomol. 143, 1-34.

Williams, L.M., Ma, X., Boyko, A.R., Bustamante, C.D., Oleksiak, M.F. 2010. SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genetics. 11, 32.

Wolfe, K.H., Sharp, P.M., Li, W.-H. 1989. Mutation rates differ among regions of the mammalian genome. Nature. 337, 283-285.

Xie, G., Chain, P.S.G., Lo, C.-C., Liu, K.-L., Gans, J., Merritt, J., Qi, F. 2010. Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. Mol. Oral. Microbiol. 25, 391-405.

Yoshido, A., Yamada, Y., Sahara, K. 2006. The W chromosome in several lepidopteran species by genomic *in situ* hybridization (GISH). J. Insect Biotechnol. Sericology. 75. 147-151

Zellmer, A.J., Hanes, M.M., Hird, S.M., Carstens, B.C. 2012. Deep phylogeographic structure and environmental differentiation in the carnivorous plant *Sarracenia alata*. Syst. Biol. 61, 763-777.

Zhao, J.L., Zhang, Y.Y., Luo, A.R., Jiang, G.F., Cameron, S.L., Zhu, C.D. 2011. The complete mitochondrial genome of Spilonota lechriaspis Meyrick (Lepidoptera: Tortricidae). Mol. Biol. Rep. 38, 3757-3764.

Zhao, J.L., Zhu, C.D. 2012. Mitochondrial Genomes of *Acleris fimbriana* (Lepidoptera: Tortricinae: Acleris) and *Hoshinoa longicellana* (Lepidoptera: Tortricinae: Acleris). HQ452340.1 Submitted to the GenBank database, December 25, 2012.

Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., Vagner, L.L., Khaspekov, G.L., Kozhemyako, V.B., Matz, M.V., Meleshkevitch, E., Moroz, L.L., Lukyanov, S.A., Shagin, D.A. 2004. Simple cDNA normalization using Kamchatka crab duplex-specific nuclease. Nucleic Acids Res. 32, e37.

Chapter 3

Diagnostic single nucleotide polymorphisms for species of the spruce budworm complex (*Choristoneura fumiferana*)

3.1. Introduction

Understanding the genetic basis for the differentiation of species provides insight into the evolution and life history differences of those species (Olson and Varki, 2003; Lee, 2002). Species differentiation is often produced by divergent selection on adaptive traits (Nosil, 2012). Genetic correlations have been found for adaptive traits such as species specific sex pheromone receptors (Wanner et al., 2010; Leary et al., 2012) and pheromone production (Lassance et al., 2010), success of invasive species (Lee, 2002), pest resistance to host toxins (Jones, 1998), social genetic responses to disease infection of neighbours (Silva et al., 2013), and susceptibility of host trees (Newton et al., 1999). Variation in adaptive traits may be measured between populations or species using candidate genes (González-Martínez et al., 2006; Hancock et al., 2008; Hufford et al., 2012), defined as genes that are hypothesized to produce a phenotype that is similar to a homologous gene in another organism (Fitzpatrick et al., 2005). Candidate gene lists are usually obtained from genome databases or known genes of interest (Fitzpatrick et al., 2005). To gain a novel set of candidate genes that distinguish populations, one approach is to randomly sample the genome for SNPs (single nucleotide polymorphisms) that show evidence of being under selection or simply distinguish a population. Then the sequences surrounding these SNPs can be annotated to infer biological function from homologous genes (Hufford et al., 2012; Bonin et al., 2009).

Candidate genes for adaptive traits involved in species differentiation can help us understand the evolutionary processes involved in closely related species complexes. The spruce budworm (SBW) species complex, *Choristoneura*

fumiferana (Clemens 1865) is an economically important forest pest group in North America (Volney and Fleming, 2007). The taxonomic history of the eight species in this complex is complicated and some species include more than one genetically distinct biotype (Freeman, 1953; Lumley and Sperling, 2011). In 1953, Freeman described the form feeding on jack pine (*Pinus banksiana*), *C. pinus* Freeman 1953, distinguishing it from the spruce/balsam feeding *C. fumiferana*. Descriptions of *C. biennis* Freeman 1967, *C. occidentalis* Freeman 1967, *C. orae* Freeman 1967, and *C. viridus* (synonymized with *C. retiniana* (Walsingham 1879) (Brown et al., 2005)) followed 14 years later (Freeman, 1967). The morphological variation, ecological characteristics, and geographic ranges of these species all partially overlap with at least one neighbour (Freeman, 1967; Lumley and Sperling, 2011). Yet, the species appear to remain differentiated despite retaining the capability to interbreed (Sanders et al., 1977).

Choristoneura pinus, the jack pine budworm, is more ecologically and morphologically distinct and has stronger support for monophyly than most SBW species (Harvey, 1996; Sperling and Hickey, 1994; Sperling and Hickey, 1995; Lumley and Sperling, 2011). Morphologically it is noticeably smaller than other SBW species, often the thorax and forewing are more reddish in colour (Lumley and Sperling, 2011), and the apex of the uncus on male genitalia is narrower than that of *C. fumiferana* (Freeman, 1953). Moths fly in July about two weeks later than *C. fumiferana* (Freeman, 1953). The main host plant of subspecies *C. pinus pinus* is jack pine and *C. pinus maritima* Freeman 1967 feeds on both *P. virginiana* and *P. rigida* (Freeman, 1967). The other coniferophagous *Choristoneura* species feed on spruce, fir, larch, Douglas fir (*Pseudotsuga menziesii*), and, in the case of *C. lambertiana* (Busck 1915) in Canada, lodgepole pine (*P. contorta*) (Lumley and Sperling, 2011; Freeman and Stehr, 1967) and several other species of pine in the USA (Powell, 1995; Powell and De Benedictis, 1995; Lumley and Sperling, 2011).

This chapter aims to determine the genetic basis of biological innovations that differentiate SBW species. To find these biological innovations we search for SNP genotypes unique to each species. Loci with a genotype, or private allele,

found only in one species or population and absent in all others, can be considered diagnostic, if specimens were sampled across its entire range and in adequate numbers. These unique SNPs can be designated as apomorphies if the genotype unique to the group is phylogenetically determined to be the derived state. If such a unique derived genotype characterizes a species, then it is an autapomorphy. Similarly, if it characterizes a clade of sister species, it is a synapomorphy. However, if the SNP genotype is not the derived state, but the ancestral state, it would be a plesiomorphy. In this chapter, unique SNPs were found for *C. biennis*, *C. carnana*, *C. fumiferana*, *C. occidentalis*, *C. pinus*, *C. retiniana*, and two *Choristoneura* species that are not in the SBW complex, *C. rosaceana* and *C. conflictana*. We focussed on *C. pinus*, because of its clear ecological distinctions from *C. fumiferana* despite extensive sympatry. The biological functions of proteins homologous to the flanking sequence of these SNPs were then compiled and evaluated for their potential to contribute to speciation.

3.2. Methods

3.2.1. Sample collection, DNA preparation, and Genotyping by Sequencing

The sequence data used in this study was collected for Chapter 2, which employed Illumina sequencing of DNA associated with restriction sites to sample SNP variation among species and specimens of the spruce budworm species complex and two outgroup species. The data included 99 specimens whose DNA was digested with the ApeKI restriction enzyme (6 SBW species and 2 outgroups, with 2 to 41 specimens per species), and 144 specimens digested with PstI-MspI (5 SBW species and 2 outgroups with 2 to 65 specimens per species) (Table 3-1). Eighteen specimens were sampled in both analyses (*C. pinus* N=4, *C. fumiferana* N=4, *C. occidentalis* N=4, *C. biennis* N=2, *C. carnana* N=1, *C. conflictana* N=1, *C. rosaceana* N=2). The last two species are deciduous tree feeders placed in *Choristoneura* (Vakenti et al., 2001; Lindroth, 1991) but not in the SBW complex, and are included here as outgroups.

3.2.2. Filtering sequences and the TASSEL pipeline

The ApeKI and the PstI-MspI sets of Illumina sequences were analyzed separately using the TASSEL pipeline (Bradbury et al., 2007), as described in Chapter 2. We ran parallel analyses using a read depth threshold of two and five reads. Alignment to the April 19, 2011 reference genome for *C. fumiferana* supplied by Roger Levesque (Institut de biologie intégrative et des systèmes, Université Laval, QC) and Michel Cusson (Natural Resources Canada, Laurentian Forestry Centre, QC) (560,420 contigs) was performed using the Burrows-Wheeler alignment tool (Li and Durbin, 2009) on default settings. Sequences that aligned to multiple positions or did not align were removed.

After searching for SNPs, SNP loci were filtered to remove gaps, loci with three or four alleles, and loci with a minor allele frequency of less than 0.02 (the setting for unrelated individuals recommended in the TASSEL documentation, Glaubitz et al, 2012). The TASSEL program recognizes SNPs based on the position of the start of the sequence read. Sequence reads that overlap on the genome can produce SNPs with duplicate entries. Duplicate SNPs were merged if they shared the same pair of alleles and if the mismatch rate for that locus was two or lower. The mismatch rate of a locus was equal to the number of specimens with genotypes that disagreed, divided by the number of specimens genotyped in both SNPs. This lenient mismatch threshold was used because heterozygotes were common (procedure recommended by the TASSEL project manager, J. Glaubitz, pers. comm. January 23, 2012). Heterozygotes were called for merged SNPs, and unmerged SNPs, which were likely paralogous loci, were removed. The merged SNP loci were filtered a final time with a minimum minor allele frequency of 0.02, and a SNP matrix was created for all SNP loci genotyped in 75% or more specimens.

3.2.3. Diagnostic SNPs, sequence grabbing, and gene annotation

The number of unique SNPs detected should depend on the extent of divergence of each species and the number of specimens being compared. A comparison among only 5 specimens would not allow detection of a SNP genotype that is truly unique to a species as reliably as a comparison between 75 specimens, because an apparently unique locus may be lost with the inclusion of additional specimens that do not share the same allele. Similarly, a locus genotype that appears to be unique for a species when compared to only 2 other species, may actually be shared when compared to additional species. To avoid false designation of unique loci, only loci genotyped in 75% or more specimens were used. False designations were avoided by removing loci with extensive missing data, but this filter was balanced by the desirability of analysing as many loci as possible to understand the overarching biological functions involved in the differentiation of these species. Comparisons to outgroup species were used to judge if unique genotypes were derived (apomorphic), and the designation of autapomorphic was only used for unique SNPs of ingroup species with larger sample sizes.

Loci with a genotype that consistently occurred in one SBW species and no other species were counted as unique SNPs (*e.g.* if all *C. pinus* had an A and no other specimens had an A, M, R, or W at that locus, Ns were ignored). Unique SNPs were counted in the same way for clades reconstructed in the most parsimonious topologies from the ApeKI or PstI-MspI SNP sets (Chapter 2). Note that these clades excluded *C. retiniana* in PstI-MspI SNP sets since no specimens were sequenced for that species. Although SNPs supporting nonmonophyletic groups existed, such as *C. pinus* + *C. fumiferana*, they were not pursued here. Fasta files were generated from 100 base pairs (bp) of reference genome sequence flanking each SNP locus (200 bp total where the SNP locus is at position 101), and from 200 bp flanking sequence (400 bp total where the SNP locus is at position 201) using a custom perl script (M. Brandão, October 4th, 2012), for all SNP loci.

Using the program BLAST2GO under the default settings (version 2.6.4; Conesa et al., 2005) the 200 bp and 400 bp sequences were used to search the GenBank nonredundant database (Benson et al., 2013) with BLASTx. The top 20 hits (*e*-value ≤ 1E-3) were recorded, using the QBlast-NCBI blast mode with a high scoring pair (HSP) length cut-off of 33 and the low complexity filter on (Conesa et al., 2005; Götz et al., 2008). The *e*-value, also known as the "Expect-value" (NCBI, 2013), represents the probability of returning a match by chance alone, so smaller *e*-values are more significant. We expect the shorter 200 bp sequences to have larger *e*-values because normally there is a greater probability of finding perfect or near perfect matches to shorter sequences. The resulting gene accessions were mapped to Gene Ontology (GO) terms under default settings in BLAST2GO, using the DBXRef Table and Gene Product Table of the GO-Database. Gene names were used to search species specific entries of the Gene Product Table of the GO-Database, and the GI identifiers were used to retrieve UniProt IDs from protein databases in PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB. These GO terms were annotated using an *e*-value hit filter of 1.0E-6, annotation cut-off of 55, GO weight of 5, and HSP-hit coverage cut-off of 0 in the subsequent BLAST2GO step (default settings). An InterPro scan (Quevillon et al., 2005) was run in BLAST2GO using all available databases (BlastProDom, FPrintScan, HMM-PIR, HMM-Pfam, HMM-Smart, HMM-Tigr, ProfileScan, PatternScan, SuperFamily, Gene3D, HMM-Panther, SignalP, and TM-HMM). InterPro GOs and annotations were merged and GO-enzyme code mapping was run using BLAST2GO under default settings.

The proportion of SNPs returning BLASTx hits and proportion of top hits with GO terms were calculated across all SNPs, unique SNPs for each species, and synapomorphic SNPs for each clade. The alignment lengths of top BLASTx hits to *C. pinus* autapomorphic SNP flanking sequences were graphed to compare the 200 bp and 400 bp sequence lengths. The *e*-values and % identity of the top hits to *C. pinus* autapomorphic SNP flanking sequences were also graphed to compare the sequence lengths. Alignment lengths, *e*-values, and % identities of

the 400 bp SNP sequences of all SNPs found in the ApeKI and PstI-MspI datasets were graphed to compare the results from the different restriction enzymes.

Gene Ontology terms are organised into categories, where the GO level 1 categories are “biological process”, “molecular function”, and “cellular component”. Inside the GO level 1 categories, there are subsequent levels of more specific categories. The GO level 2 categories inside biological processes, molecular functions, and cellular components were graphed for top hits to *C. pinus* autapomorphic SNP flanking sequences, all ApeKI SNP flanking sequences, and all PstI-MspI SNP flanking sequences. The more specific GO level 3 categories were also graphed for top hits to *C. pinus* autapomorphic SNP flanking sequences, for within each of the broad GO level 1 categories, biological processes, molecular functions, and cellular components. The sequence descriptions from the top hits to the *C. pinus* autapomorphic SNP flanking sequences were organised by general function and tabulated.

3.3. Results

3.3.1. Descriptive Genotyping by Sequencing results

The ApeKI restriction enzyme was predicted to have roughly 7 times more restriction sites than PstI-MspI (Table 2-1 of Chapter 2). Unique ApeKI sequences with 2 or more reads did have a lower read depth (averaging 27.8 reads per allele compared to PstI-MspI’s 49 reads per allele). ApeKI also produced roughly 4 times more SNP loci than PstI-MspI, and lower genotyping coverage of specimens (overall average 0.415 for ApeKI, and 0.514 for PstI-MspI, with a locus genotype coverage threshold of 10%) (Fig. 2-6 of Chapter 2). Since the recognition sequence of ApeKI is shorter than that of PstI, the greater number of ApeKI sites was expected, which in turn contributed to the lower read depth and genotyping coverage of ApeKI. In addition, nearly all PstI-MspI reads contained a barcode and cut site (0.967), whereas less than half of the ApeKI reads did

(0.477), possibly because the initial ApeKI sequencing runs were tests of a newly set-up library preparation and sequencing platform.

Choristoneura fumiferana, the species used as reference genome, averaged the highest genotyping coverage, and other species decreased generally in order of taxonomic distance (western species, *C. pinus*, outgroups), presumably because of an increased frequency of mutations in the restriction enzyme recognition sequence of species more diverged from *C. fumiferana* (Chapter 2). Maximum parsimony, maximum likelihood, and neighbour-joining phylogenies of the SNP data all placed *C. pinus* as sister to *C. fumiferana* and the western species (Figs. 2-8, and 2-9 of Chapter 2). This is a novel topology that is incongruent with previous mtDNA-based phylogenies (Sperling and Hickey, 1994; Sperling and Hickey, 1995; Lumley and Sperling, 2011). Further description of these phylogenetic results is found in Chapter 2 of this thesis.

3.3.2. Diagnostic SNPs and BLASTx search results

The read depth threshold of two reads per allele provided more data than the threshold of five because the read depth filter removes data early in the pipeline before duplicate SNPs are merged. In a low coverage dataset like ApeKI, raising the read depth threshold could increase false homozygous base calls (because the alternate allele is removed). Therefore, in order to gain the most complete coverage of genomic variation, a read depth threshold of two was used in all further analyses.

The numbers of unique or autapomorphic SNPs for each species are described at greater length in Chapter 2 but are briefly reiterated here (Table 3-1). *Choristoneura pinus* produced the largest number of autapomorphic SNPs of the coniferophagous *Choristoneura* species in the datasets for both ApeKI (n=8, SNPs=594) and PstI-MspI (n=65, SNPs=353). In this species, and also the outgroup species, which scored an even larger number of unique SNPs, some of the same loci were sequenced in both restriction enzyme analyses, so the total number of loci was less than the sum of the two sets (Table 3-1). It is important to

note that the number of unique SNP genotypes found for a species/cluster was heavily influenced by the number of specimens (*i.e.* it was "easier" to find private alleles for only 2 individuals as opposed to 45).

Increasing the sequence length from 100 bp of sequence flanking each side of the SNP to 200 bp increased the number of BLASTx top hits to these flanking sequences from 198 to 278 for the 945 unique SNP loci in *C. pinus* (21.0% to 29.4%) (Table 3-1, and Fig. 3-1), and decreased the average *e*-value (1.58E-5 in APM200pi to 1.48E-5 in APM400pi) (Fig. 3-2). However, increasing the query length from 200 to 400 bp also increased the number of accepted short alignments that ceased prior to the SNP position, rather than including the SNP, from 0 (all >34aa alignment length) to 140 (<67aa alignment length) in *C. pinus* (Fig. 3-1). Increasing the query sequence length also decreased the average sequence similarity (83.11% to 80.14%) (Fig. 3-3). Recent studies on maize evolution considered genes within a 10kb window to be linked to their SNPs of interest (Hufford et al., 2012). Our 400 bp window is conservative in comparison, and because we were interested in SNPs within coding sequence, as well as closely linked regions, all results from the 400 bp sequences were analysed further.

The percentage of SNP flanking sequences with protein matches varied between species and restriction enzymes (Table 3-1). SNP flanking sequences from ApeKI recovered more matches than PstI-MspI when considering more distinct species (high numbers of unique SNPs) or groups with larger numbers of specimens. There is a large difference between restriction enzymes in the case of *C. conflictana*. Only 2 specimens were sequenced for *C. conflictana* in PstI-MspI, so a genotype unique for this species would have an allele frequency of about 0.01 if all 144 specimens were genotyped at the locus. The minimum minor allele frequency filter was set to 0.02, so the majority of these SNPs were filtered out. However, not all SNPs were genotyped in all 144 specimens. The locus genotype coverage filter was 75%, so there were some loci genotyped in only 108 specimens. Some of these loci were diagnostic for *C. conflictana* and the allele frequency would be about 0.02. So, although we expected *C. conflictana* to have many diagnostic loci, which they did from ApeKI sequences, there were only 104

loci from PstI-MspI because it was constrained by the minimum minor allele frequency filter (Table 3-1).

Between species, the *C. carnana* SNP sequences recovered the fewest BLASTx matches proportionally (23.3%) and *C. conflictana* the most (64.3%) (Table 3-1). There was a trend that SNP sequences from more divergent species (*e.g.* outgroup *C. conflictana*) tend to recover a higher percentage of hits. When all SNP flanking sequences were searched, regardless of being diagnostic or not, 38.5% recovered BLASTx hits (Table 3-1).

Both ApeKI and PstI-MspI produced a very similar distribution of alignment lengths for BLASTx top hits (Fig. 3-4), except for the shortest alignments where PstI-MspI had proportionally more than ApeKI. It could be that the broader sampling of the genome by ApeKI increased the quality of its distribution. There were no alignments less than 33 aa in length because of the high scoring pair (HSP) filter. There was an upward slope peaking in the 50-60 aa bin, because longer alignments produce higher scores and are more likely to be a top hit. However, this begins to decrease in the 60-70 aa bin where alignment lengths begin to include the SNP loci, and it appears that SNPs are more likely to be found outside or near the end of conserved coding sequences. There was a final peak in the 130-140 aa bin, representing full alignments to the querying 133.3 aa sequence, and the top hits beyond this maximum length were from alignments that included gaps. The percent identity and *e*-value distributions were very similar for ApeKI and PstI-MspI (Figs. 3-5, and 3-6).

The pattern of alignment lengths to the 400 bp *C. pinus* SNP sequences (Fig. 3-1) is similar to the pattern seen in all SNPs (Fig. 3-4). Similarly, the alignment lengths to the 200 bp *C. pinus* SNP sequences is a compressed version of this pattern (Fig. 3-1).

3.3.3. Gene Ontology mapping results

We focused on the GO results of *C. pinus*, as opposed to the other SBW species, because its greater number of diagnostic SNPs provided an opportunity to

consider proportions of functions of the homologous sequences. About 64% of *C. pinus* genes returned GO term assignments (Table 3-2), which is higher than *C. fumiferana* but lower than most of the other species. Gene Ontology terms were scored by the number of sequences supporting that term, and also their proximity to it (Conesa et al., 2005). In this way the broadest categories often score highest, but the highest scoring category did not always have the most sequences supporting it. The top scoring biological processes were cellular process and metabolic process (Table 3-2), and the top scoring molecular functions were binding and nucleic acid binding (Table 3-2) for species and species clusters with a large number of apomorphic SNPs (Table 3-1).

The biological processes supported by the most *C. pinus* sequences were cellular and metabolic processes (Fig. 3-7), and the most common molecular functions were binding and catalytic activity (Fig. 3-8). The cellular components where the highest numbers of these processes and functions take place were the cell and organelle for *C. pinus* (Fig. 3-9). These were also the most common biological processes, molecular functions, and cellular processes when considering all SNPs found in ApeKI and PstI-MspI associated sequences.

The *C. pinus* autapomorphic SNP flanking sequences were homologous to 34 top hit sequences involved in detoxification and immune response, 13 in metabolism, 16 in sensory perception and motor control, 9 in morphology, 17 in cell cycle, 3 in cell movement, 48 in intracellular and cell-cell communication, 54 in gene expression, 33 in mobile genetic elements, 9 in multiple functions, and 42 in unknown functions (Table 3-3). The majority of these top hit sequences were from *Danaus plexippus*, the monarch butterfly.

3.4. Discussion

3.4.1. Comparison to other species

Larger numbers of unique SNPs were found in species either because the species was evolutionarily divergent, or because there were few specimens

representing that species, leading to more shared SNPs simply by chance. A high number of unique SNPs were found for *C. retiniana*, probably not because it is a more divergent species (it is phylogenetically placed within the western lineage), but because it had few specimens (Table 3-1). However, the reason why *C. pinus*, which had few specimens in the ApeKI analysis but many in the PstI-MspI analysis, had a large number of SNPs was probably because it was more diverged.

Proportionally more BLASTx hits were returned for sequences containing SNPs distinguishing the outgroup species (*C. rosaceana* and *C. conflictana*) and ingroup clade from each other, than for SNPs distinguishing single species in the SBW complex (Table 3-1). These were comparisons between more diverged groups. More distantly related species share fewer restriction enzyme recognition sites (Chapter 2; Arnold et al, 2013). However, the few recognition sites shared are more likely to be in conserved areas of the genome, because conserved areas are less likely to retain mutations (Dale et al., 2012). Conserved areas usually are functional or coding sequence (Dale et al., 2012), which would explain why more BLASTx hits are returned. Alternately, a higher proportion of BLASTx hits could be due to a randomly higher similarity between some areas of the *C. fumiferana* reference genome and the sequences in the non-redundant NCBI sequence database. However, a randomly uneven similarity is unlikely to produce the pattern of increasing proportions of BLASTx hits as the divergence between groups increases. As a measure of divergence, this pattern means *C. pinus* is slightly more diverged than the other SBW species from the *C. fumiferana* reference genome, because it returned BLASTx hits for about 30% of sequences as opposed to about 25% of sequences (Table 3-1).

When all SNP flanking sequences were used as queries, the percentage that returned BLASTx hits, 38.5%, was comparable to the results of transcriptome annotation studies (Bissinger et al., 2011; Vogel et al., 2011), although we expected the majority of SNPs to be in non-coding regions. The proportion of BLASTx hits is likely increased by the SNPs proximity to restriction enzyme recognition sites conserved between specimens. A tick transcriptome recovered matches for 28.6% of their sequences (Bissinger et al., 2011), and a greater wax

moth EST (expressed sequence tag) library recovered 40% (Vogel et al., 2011). The 64.2% of *C. pinus* gene matches with GO term assignments was also comparable to results in the literature, where 45% greater wax moth genes (Vogel et al., 2011), 72.2% tick genes (Bissinger et al., 2011), and roughly 46.3% *Drosophila* genes in an entire genome annotation study (Adams et al., 2000) recovered GO term assignments.

The categories of biological processes and molecular functions containing the most *C. pinus* sequences were representative of the SNP variation as a whole, as seen in ApeKI and PstI-MspI (Figs. 3-7, 3-8, and 3-9, and Table 3-2). ApeKI and PstI-MspI sequences returned almost identical proportions of biological processes and molecular functions, which means that the variation is distributed evenly regardless of enzyme (Figs. 3-7, 3-8, and 3-9, and Table 3-2). The categories found in ApeKI, PstI-MspI, and *C. pinus* were comparable to results found in whole genome or transcriptome studies (Vogel et al., 2011; Bissinger et al., 2011; Adams et al., 2000), indicating that our results were a good representation of the genes present in the *Choristoneura* genome. This representation also means we included enough loci in our apomorphic SNP search, and that the 75% genotype coverage threshold was adequate to sample the entire genome for SNPs involved in differentiation of the major SBW lineages. The two Gene Ontology level 2 (the broadest category) biological processes with the most sequences of the tick transcriptome (Bissinger et al., 2011) were the same in ApeKI, PstI-MspI, and *C. pinus*; cellular process and metabolic process (Fig. 3-7). However, when comparing categories with fewer sequences supporting them the similarity of the tick transcriptome to *C. pinus* decreases (Bissinger et al., 2011). In the tick the biological processes GO level 2 categories with the next most sequences were localization, biological regulation, and cellular component organization (Bissinger et al., 2011), whereas *C. pinus* had biological regulation, response to stimulus, and signaling (Fig. 3-7).

On level 3 the three biological processes with the most sequences in both *C. pinus* (Fig. 3-10) and the greater wax moth were the same; primary metabolic process, cellular metabolic process, and macromolecule metabolic process (Vogel

et al., 2011). Again, the similarity decreases in the categories with fewer sequences, where the next biological process on level 3 for *C. pinus* was nitrogen compound metabolic process (Fig. 3-10) which did not appear on the list in the greater wax moth transcriptome (Vogel et al., 2011). Most major categories of the level 3 molecular functions represented by *C. pinus* sequences were similar to those of the greater wax moth transcriptome, namely hydrolase binding, nucleic acid binding, protein binding, nucleotide binding, and transferase activity, although the proportions were different and ion binding, a well-represented category in *C. pinus*, was rare in the greater wax moth (Vogel et al., 2011; Fig. 3-10).

3.4.2. Sequences potentially implicated in speciation of C. pinus

Broad categories of GO terms indicate coverage of the genome, however to understand the biological innovations associated with speciation, we examine the functions of individual genes. *Choristoneura pinus* produced the largest number of autapomorphic SNPs in the SBW group (Table 3-1) and returned a higher proportion of BLASTx hits than most SBW species, indicating that this species was more distantly related to the other species in the SBW group. It could be evidence of an earlier divergence time, or of more rapid evolution and divergent selection, which makes the *C. pinus* SNPs an interesting case study.

In the following sections we describe the gene products associated with the unique and presumably autapomorphic SNPs of *C. pinus* and the highlights of the major functional groups. It is important to remember that half of these sequences were linked (140/278) and half contained the SNP in the coding sequence. Of the sequences containing the SNP in the coding region, some will be synonymous, and some will be non-synonymous. Synonymous SNPs should not affect the function of the protein. Even non-synonymous SNPs, if located in an amino acid not involved in the folding of the protein or the active site, may not affect the function of the protein. If a non-synonymous SNP changes the amino

acid to one with a similar charge or binding properties, it also may not affect the function of the protein.

It is also important to remember that the average % identity of the sequence segment aligned was about 80%. Although these sequences were homologous to the *C. pinus* sequences, even small changes in the sequence can change the function of the protein. To ascertain the function of these proteins in *C. pinus*, mutation or knock-out studies will need to be performed. The following list describes the functions of homologous proteins found in other insects, and acts as an "aerial view" of landmarks in the SBW genome that could be involved in adaptive speciation.

3.4.2.1. Detoxification and immune response

One of the major life history differences between *C. pinus* and the other SBW species is its ability as larvae to digest pine instead of spruce or fir (Freeman and Stehr, 1967). Trees produce terpenoid compounds to deter herbivores (Wallin and Raffa, 1999; Ralph et al., 2006) and the herbivores deactivate these compounds with detoxifying enzymes (Scott et al., 1998). Oxidation-reduction enzymes are known to be involved in detoxifying plant compounds, and there are a number of oxidation-reduction genes in the *C. pinus* SNP list. Cytochrome P450, aael004336-partial protein, sulfide:quinone mitochondrial-like protein, the breast cancer metastasis-suppressor 1-like protein, prophenoloxidase 1, wd-repeat protein, glucose dehydrogenase, and 3-hydroxyacyl- dehydrogenase type-2 have oxidoreductase activity (Table 3-3). Some, like cytochrome P450, for which there were three SNPs fixed for *C. pinus*, are known to catalyze the oxidation of organic substances including lipids, hormones, and toxins and specifically terpenoid compounds (Schuler, 1996; Scott and Wen, 2001; Scott et al., 1998).

Among the other detoxifying enzymes with SNPs fixed for *C. pinus* was the major allergen protein. This protein is known to be involved in insect host plant interactions (Fischer et al., 2008). The major allergen protein belongs to a fast-evolving gene family, where gene duplications were found by Fischer et al.

(2008) to be a driving force of speciation and adaptation in some lepidopteran species (Pieridae).

Esterases modify specific proteins and are involved in the insect midgut detoxification system (Schuler, 1996); fixed SNPs for *C. pinus* were found in carboxylesterase and ubiquitin thioesterase. *Choristoneura pinus* SNPs were also found in ubiquitin-protein ligases which attach ubiquitin to a lysine of protein substrates to target it for degradation by the proteasome. Fixed SNPs for *C. pinus* were found in the vitamin k-dependent protein c which is involved in proteolysis, and dipeptidase which is excreted by the midgut and degrades proteins (Zhu et al., 2011).

The *C. pinus* list contains proteins involved in the immune system including tyrosine-protein kinase csk which suppresses signalling by the T-cell receptor (Chow et al., 1993), basigin which belongs to the immunoglobulin superfamily (Igakura et al., 1996), activating transcription factor- isoform b which is involved in response to salt stress (Sano et al., 2005), leucine-rich repeats and immunoglobulin-like domains protein 2, heat shock protein 90 beta, and c-maf-inducing protein which is involved in T-cell signalling. The immune system targets and removes or degrades foreign compounds which could be important for the host specificity of *C. pinus*.

3.4.2.2. Metabolism

Similar to the way in which detoxification degrades foreign substrates, metabolism regulates the degradation of glucose, lipids, carbohydrates, and regulates cholesterol levels. There were *C. pinus* SNPs in krueppel-like factor 15 (KLF15), which regulates the transcription of other genes including Wnt/β-catenin that affect heart cells (Noack et al., 2012), and regulates glucose and amino acid metabolism and skeletal muscle lipid utilization (Haldar et al., 2012). The wd repeat and fyve domain-containing protein 3-like is involved in the carbohydrate metabolic process, and lipase 3-like metabolizes lipids. The tafazzin homolog, found in *C. pinus*, metabolizes cardiolipin in muscles (Xu et al., 2006),

and 3-hydroxyacyl- dehydrogenase type-2 is involved in fatty-acid metabolism. Insulin-like growth factor 2 mRNA binding protein regulates RNA translation essential for growth and development (Christiansen et al., 2009). High density lipoprotien binding protein vigilin binds high density lipoprotein which could regulate excess cholesterol levels in cells (Chui et al., 1997). Guanylate cyclase 32e-like, luciferase, and abhydrolase domain containing 11 are similarly involved in metabolism.

The host plant differences could explain SNPs found in genes influencing the metabolism of glucose, lipids and carbohydrates. Metabolism also influences energy allocation (Izadi et al., 2011), ability to overwinter (Han and Bauce, 1998; Pullin, 1987; Chown and Gaston, 1999), timing of emergence (Graham et al., 1980), egg production (Ishihara and Shimada, 1995), and other life history traits that are known species differences in the SBW complex.

3.4.2.3. Circadian clock, flight, and sensory perception

There were two *C. pinus* genes homologous to genes that influence insect circadian clock and flight times. The nuclear receptor subfamily 2 group f member 6 is a steroid hormone receptor and transcription factor. There is some evidence that this gene effects the entrainment of the circadian clock by photoperiod from mutant phenotype experiments done in mice (Warnecke et al., 2005). Secondly, experiments with the gene *shaggy* in *Drosophila* demonstrate that it affects their circadian clock and flight time (Martinek et al., 2001). There are species differences in the time that female SBW moths call in the evening (Sanders et al., 1977). The hour when 50% of female *C. pinus* call is four hours later that of female *C. fumiferana* (Sanders et al., 1977), so these genes could be influencing this species difference.

There are a few genes in the *C. pinus* list that influence motor control. The turtle protein is essential in the establishment of coordinated motor control, and *Drosophila* with a mutated form of this gene were unable to turn over and remained on their backs like turtles (Bodily et al., 2001). Protein stoned-a

influences flight behaviour (Homyk and Sheppard, 1977), and the potassium voltage-gated channel protein eag (Griffith et al., 1994) influences flight behaviour and mating. The beta adrenergic-like octopamine receptor influences motor ability (Wu et al., 2012), the putative otopetrin gene belongs to a family of genes known to influence spatial orientation and acceleration (Hughes et al., 2004). Divergent sexual selection often acts on mating behaviour to produce ecological speciation (Nosil, 2012), so many of these genes could be under selection cumulating in *Choristoneura* species differences.

Sensory perception is important for finding and attracting mates, and differing abilities to perceive mate characteristics can lead to divergent selection (Nosil, 2012). *Choristoneura pinus* has a fixed SNP in semaphorin 2a, which influences salivary gland development, drinking behaviour, visual behaviour, and flight behaviour. Serine threonine-protein kinase doa influences vision and sensory transduction (Yun et al., 1994), and, similarly, otoferlin-like protein is involved with hearing and sensory transduction. The ankyrin repeat domain-containing protein has calcium channel activity and is involved in the sensory perception of sound and mechanosensory behaviour. The intraflagellar transport protein 140 homolog also is involved in the sensory perception of sound, and nonmotile primary cilium assembly.

Our one pine feeding species, *C. pinus*, had a fixed SNP in the gustatory receptor 45, which is a taste receptor. Gustatory receptors in *Heliconius* butterflies have been shown to be more strongly expressed in females than males and play a role in selecting a palatable host plant on which to oviposit (Briscoe et al., 2013). It is possible that it plays a role in host plant differentiation in SBW species. And finally, and possibly the most interestingly, there is a match in the *C. pinus* list to an odorant receptor sequence found in the codling moth *Cydia pomonella* (Bengtsson et al., 2012), which is another tortricid, a close relative of *Choristoneura*. Spruce budworm moths attract mates using species-specific pheromones (Silk and Kuenen, 1988), so this gene is a key candidate for influencing a species defining trait.

3.4.2.4. Morphology

We found genes involved in wing disc development in the *C. pinus* list, and because they are morphologically smaller and more reddish than the other species (Freeman, 1967) these genes could contribute to these species differences. Wing pattern is one of Nosil's "magic genes" (2012) which are involved in species divergence in some Lepidopteran species (Fordyce et al., 2002). Mutant experiments on the scabrous protein in *Drosophila* produced incomplete wing margin development (Lee et al., 2000), signifying that this protein is involved in wing development. We found matches for the trithorax group protein osa on two different contigs, suggesting paralogs of this gene. It is similarly involved in wing disc dorsal/ventral pattern formation (Terriente-Félix and de Celis, 2009), along with oligosaccharyl transferase which influences chaeta development.

Body morphology is another trait that differentiates species (Nosil, 2012; Freeman, 1967). There was a *C. pinus* SNP in the ankyrin repeat domain-containing protein 12 (a gene similar to the one involved in the perception of sound above, but on a different contig) which influences head morphogenesis. Nonclathrin coat protein gamma1-cop is involved in the formation of the tracheal system, and both it and the hormone receptor-like in isoform d are involved in cuticle development. The glycylpeptide n-tetradecanoyltransferase 2 protein is involved the dorsal closure of the embryo.

Melanisation is often a visible effect of evolution and speciation (True, 2003), and there are two genes with *C. pinus* SNPs involved in this. Prophenoloxidase 1 which is involved in melanin biosynthesis from tyrosine, and heat shock protein 90 beta which is in the melanosome. Both these genes also have functions in detoxification and immune response (Table 3-3).

3.4.2.5. Cell cycle, mitotic spindles, cell adhesion, and fertilization

Development of organisms into different forms is influenced by cell proliferation, so autapomorphic SNPs in genes influencing the cell cycle can be

expected. The huntingtin interacting protein 1 which contains a death effector domain, cell cycle checkpoint kinase 2, myotubularin-related protein 9-like, and ww domain-containing oxidoreductase (Chang et al., 2005) are involved in apoptosis control. Cdc2-related kinase (Meyerson et al., 1992), cell division cycle protein 23 homolog, and cyclin g are involved in the cell cycle. The nipped-b-like protein and three hypothetical proteins from *Danaus plexippus* are putatively involved in chromatin binding, which is important during cell division. Spliceosomal protein sap repairs damaged DNA and is involved in nuclear mRNA splicing by assembling the spliceosome (Das et al., 1999). Transcription factor rsv1 and sin3a-associated protein sap130 (Fleischer et al., 2003) are sequence specific transcription inhibitor during mitotic division. Both sap proteins are involved in mitotic spindle organization; spindles bind and organize chromosomes when the cell splits during mitosis.

There were two SNPs fixed in *C. pinus* genes which influence cell adhesion. The cd9 antigen is important for cell adhesion of the egg and sperm during fertilization (Higginbottom et al., 2003), and the insulin-like growth factor-binding protein complex acid labile chain is involved in protein-protein interaction of cell adhesion.

3.4.2.6. Cell movement, spindles and microtubules

Species differences may arise from divergent selection on recognition proteins of organelles transported, centrosome assembly, movement of cilia, flagellar movement of the sperm. *Choristoneura pinus* SNPs were found in the axonemal dynein heavy chain which is involved in microtubule motor activity and occurs in the cilium axoneme, the probable e3 ubiquitin-protein ligase mycbp2 which is found in the microtubule cytoskeleton or axon and is involved in branchiomotor neuron axon guidance, and the ect2-like isoform 3 which is involved in cytokinesis.

3.4.2.7. Cell-cell communication, action potentials, signal transduction, and transmembrane transportation

Intra and intercellular signal transduction determine how cells, and the organism as a whole, react and interact with their environment and the compounds they encounter. This involves many traits important for speciation, including organism response to toxins, odors, light, and other organisms (Nosil, 2012). It includes the transmembrane transportation of sugars, ions, and proteins, and regulates the production and reception of hormones. Roughly a quarter of the *C. pinus* SNP sequences characterized function in signal transduction, suggesting it has a major role in producing species differences.

Sequences involved in signal transduction that also matched to a biological process already described (such as sensory perception) were not recorded here. The sequences described here were annotated with general biological functions such as cell surface receptor, ligand-gated ion channel, or hedgehog receptor activity. Unfortunately more specific functions like "binds to pheromone and produces neurological signal for organism to fly towards mate" were not available.

Sequences involved in neurons included agrin-like isoform 1 and putative agrin which are involved in neuron growth (McMahan et al., 1992), and inactive ubiquitin carboxyl-terminal hydrolase 54 which is associated with neurons (Doran et al., 1983). Sequences involved in action potentials ion binding included neuropeptide receptor a33 which performs synaptic transmission in the plasma membrane for the tachykinin receptor signaling pathway, chloride transporter like-1 which transports chlorine across membranes and could be important for maintaining polarization between action potentials in muscle cells (Yuan et al., 2004), and glutamate receptor 1-like which is a neurotransmitter which has extracellular-glutamate-gated ion channel activity in the postsynaptic membrane. Sequences involved in ion binding and transmembrane transport included solute carrier family 12 member 6 which is a potassium chloride transporter, ionotropic receptor isoform f which forms an ion channel pore, and rab3 GTPase-activating

protein catalytic subunit helps regulate exocytosis of hormones and neurotransmitters.

Among the proteins involved in hormone signalling pathways, *C. pinus* has fixed SNPs in the follistatin-related protein 1-partial, steroid receptor-interacting snf2 domain protein, ultraspiracle protein, and a seminal fluid protein. Seminal fluid proteins may be subjected to adaptive selection because they influence the reproductive success of both males and females in insects (Chapman and Davies, 2004), and can be involved in species differentiation (Larson et al., 2013).

Transmembrane transport proteins included adenylate cyclase which converts ATP to cAMP, monocarboxylate transporter-like protein, proton-coupled folate transporter-like, and sugar transporter erd6-like 6. Proteins with receptor activity (bind a specific protein and transduce a signal) included low-density lipoprotein receptor, protein lin-7 homolog b-like, protein patched, protein vac14 homolog, and soluble guanylyl cyclase alpha-1 subunit which is involved in the cGMP biosynthetic process.

Jack pine budworm autapomorphies in sequences for proteins involved in intracellular protein transport included signal transducing adapter molecule 1-like, and a few proteins involved in vesicles including coatomer subunit beta, golgi reassembly stacking protein 2, and rab6 GTPase activating GAPCenA.

3.4.2.8. DNA binding and gene expression control

About a quarter of the proteins in the *C. pinus* list were either DNA or RNA binding, or were involved in sequence specific gene expression. Unfortunately, most of their annotations did not specify what sequence they were binding or the function thereof.

An example of one of the *C. pinus* sequences involved in gene expression control was Arnt (AhR nuclear translocator); Arnt is a cofactor that promotes DNA binding by the dioxin receptor (AhR) (Whitelaw et al., 1993; Lindebro et al., 1995). They belong to a super family of transcription factors, the members of

which dimerize and activate the expression of their target gene by binding upstream or downstream of the gene at its cognate binding site (Mimura et al., 1999). Arnt dimerizes with the dioxin receptor which is a ligand-activated transcription factor, and the dimer induces the transcription of enzymes which metabolize xenobiotics (environmental pollutants like 2333738-tetrachlorodibenzo-*p*-dioxin) (Mimura et al., 1999).

3.4.2.9. Mobile genetic elements

Among the *C. pinus* SNP sequences there were 20 endonuclease-reverse transcriptases, 4 reverse transcriptases, 1 retrotransposon, 1 transposon, 1 transposase, 5 proteins involved in RNA dependant DNA replication (retrotranscription), and 1 protein found in the cell envelope of gram negative bacteria (putative poll protein). Retrotransposons are class I transposable elements that make a RNA copy of themselves, reverse transcribe it into DNA, and insert it in another location in the genome. Long terminal repeats (LTRs) and long interspersed elements (LINEs) are types of retrotransposons which include the sequence for a reverse transcriptase, and are numerous in the *Bombyx mori* W chromosome (Abe et al., 2005; Kawaoka et al., 2011). It is possible that some of the SBW reverse transcriptases were native to the budworm and maintain telomeres, but it is also possible that they were incorporated into the budworm genome by horizontal transfer from viruses, parasitoids, symbiotic bacterium, or other budworms. Horizontal gene transfer increases the genomic variation of species by introducing new gene variants which can allow them to adapt to new environments and forms (Schaack et al., 2010). These proteins may have played a role in the diversification of the SBW species.

3.4.3. Species concepts and adaptive traits

Accurately defining species is important for the study of systematics, measuring biodiversity, conservation practices, and implementing successful pest

control programs (Cracraft, 2000). Species concepts were traditionally based on phenotypes (morphology, life history, phenology, etc.) (Mayr, 1982), but, with the advent of molecular technology (isozymes, protein sequences, DNA sequences, etc.), evolution based species concepts became more prominent (Yoon, 2010). Many species concepts exist, each with their own criteria. The Biological Species Concept is defined by reproductive isolation (Mayr, 1942; Mayr, 2000), the Hennigian Species Concept adds that the ancestor must cease to exist when the lineage splits (Hennig, 1966; Meier and Willmann, 2000), the Phylogenetic Species Concept can be defined by monophyly (Mishler and Theriot, 2000) or by diagnosability (Wheeler and Platnick, 2000), the Evolutionary Species Concept is defined by an independent historical fate (Simpson, 1951; Wiley and Mayden, 2000), and the Ecological Species Concept is defined by unique adaptive niche (Van Valen, 1976; Andersson, 1990). However, de Queiroz (2007; 2011) advocates a unified species concept which uses the criteria of these species concepts as lines of evidence. This species concept is an iterative approach where the more criteria a species satisfies, the more robust it is. By combining multiple lines of evidence the weakness of any one approach can be avoided (Meier, 2008).

The SBW species fill some of these criteria but not all. Geographic ranges are one of their more useful diagnostic characteristics, however the ranges of each species overlap at least partially with at least one other species (Lumley and Sperling, 2011). Host preference and morphology are often diagnostic, and accessible when identifying a moth-in-hand, but again there is some overlap between the species (Lumley and Sperling, 2010; Lumley and Sperling 2011). Biological criteria including reproductive isolation, mate recognition such as flight time, pheromones, and attractiveness (Sanders et al., 1977; Silk and Kuenen, 1988), are satisfied when ignoring small amounts of introgression or overlap, for *C. fumiferana*, *C. retiniana*, *C. pinus*, and *C. carnana*, but not between *C. occidentalis* and *C. biennis*. Similarly, phylogenetic criteria (monophyly, and fixed differences) are satisfied for all species except between *C. occidentalis* and *C. biennis* (see Chapter 2). By combining ecological and phylogenetic lines of evidence (Sanders et al., 1977; Silk and Kuenen, 1988;

Chapter 2), it becomes clear that *C. fumiferana*, *C. retiniana*, *C. pinus*, and *C. carnana* are strong and distinct species, but there is some ambiguity regarding the distinction between *C. occidentalis* and *C. biennis* (Chapter 2).

The number of unique SNPs increases in more phylogenetically distinct species (Table 3-1; Chapter 2). Many of these autapomorphies could have become fixed during an adaptive breakthrough triggering an adaptive sweep, or simply by drift and neutral bottlenecks. These SNPs not only support the species status of some SBWs and raise questions about the status of others, but also lend insight to the genetic basis of a species' integrity. The sub-set of diagnostic SNPs that influence gene function could be the biological innovations that allow SBW species to thrive in their own ecological niche and retain their ecological distinctness. These SNPs could be maintained by divergent selection, despite the movement of other genes by introgression, and produce the genomic integrity of the species, as per the cohesion (Templeton, 1989) or genomic integrity species concept (Sperling, 2003). These adaptive genes are the building blocks that diversifying selection could have acted on, and through them we may find the causes that underlie the origins and diversity of species.

3.4.4. Relevance to modern technology and future research

Candidate genes for SNP assays, gene knock-out studies, or genetic variation studies, can be efficiently selected from unique or autapomorphic SNPs mined from genotyping by sequencing data. The search parameters for apomorphic SNPs could be relaxed to 90 or 95% instead of 100% to allow for natural intraspecific variation. This could increase the number of SNPs found, especially for hybridizing, introgressing, or very closely related species.

Future research should incorporate *C. lambertiana* specimens. This is the other pine-feeding species in the complex, which we unfortunately did not have enough DNA for. By including sequence data from this species, we could understand its evolutionary relationship to the other species, especially *C. pinus*. *Choristoneura pinus* had a relatively large number of private alleles, and many of

these might belong to both *C. pinus* and *C. lambertiana*, especially if they are sister taxa. There are many questions regarding genes involved in host plant preference that further research can answer.

Because the SBW is an economically significant forest pest (Volney and Fleming, 2007), it is important to monitor the evolutionary status of its species. Large amounts of hybridization with *C. fumiferana* or switching to new hosts could be problematic for the coniferous forests of Canada. Genotyping the diagnostic SNPs found in this study in future generations of budworms would be useful for tracking the evolution of this forest pest.

3.4.5. Conclusions

In this study we identified 115,249 SNPs from Illumina sequences, of which 945 were unique or autapomorphic for *C. pinus*, making it the most genetically distinct species of the SBW complex. The GO annotation of sequence flanking these SNPs recovered biologically important genes involved in detoxification, flight, sensory perception, and morphological differences. By identifying genes associated with species differences we gain a better understanding of which genes promote the genomic integrity of species and diversifying speciation. Our results can facilitate the production of candidate gene lists and evaluation of the evolutionary significance of these loci.

Table 3-1

Numbers of samples, diagnostic SNPs, and BLASTx results, partitioned by species or clades, and associated restriction sites.

Species or clade[i]	# samples[ii]			# diagnostic SNPs			# with BLAST hit			% with BLAST hit		
	A	PM	APM	A	PM	APM[iii]	A	PM	APM[iii]	A	PM	APM
bi	8	6	14	0	4	4	0	2	2	0.0	50.0	50.0
ca	8	1	9	1	42	43	0	10	10	0.0	23.8	23.3
oc	26	8	34	0	4	4	0	1	1	0.0	25.0	25.0
oc-bi-ca	42	15	57	16	51	67	6	12	18	37.5	23.5	26.9
re[iv]	3			182			73			40.1		
oc-bi-ca-re[iv]	45			11			4			36.4		
fu	41	58[vi]	99	34	37	71	12	6	18	35.3	16.2	25.4
oc-bi-ca-re-fu[iv]	86			315			90			28.6		
oc-bi-ca-fu[v]		73			56			12			21.4	
pi	8	65	73	594	353	945	197	82	278	33.2	23.2	29.4
oc-bi-ca-re-fu-pi[iv]	94			6840			4126			60.3		
oc-bi-ca-fu-pi[v]		138			117			66			56.4	
ro	3	3	6	4697	2245	6840	2937	1184	4121	62.5	52.7	60.2
co	2	2	4	3830	104[vii]	3931	2448	81	2529	63.9	77.9	64.3
All	99	144	243	57418	59380	115249	26592	18637	44336	46.3	31.4	38.5

[i]Species are abbreviated using the first two letters: bi = *C. biennis*, ca = *C. carnana*, oc = *C. occidentalis*, re = *C. retiniana*, fu = *C. fumiferana*, pi = *C. pinus*, ro = *C. rosaceana*, co = *C. conflictana*.

[ii]Samples not specimens, as some specimens were used in both ApeKI (A) and PstI-MspI (PM) and so were sampled twice.

[iii]Combined number of SNPs and blast hits may be less than the sum due to identical SNPs from ApeKI and PstI-MspI; BLASTx searches were conducted on 400 bp query sequences.

[iv]For ApeKI only, as there were no *C. retiniana* specimens sequenced with PstI-MspI.

[v]For PstI-MspI only, because the equivalent clade in ApeKI includes *C. retiniana*.

[vi]Hybrid specimen not included.

[vii]Artificially decreased because the number of specimens (2/144) was lower than the minimum minor allele frequency (0.02).

Table 3-2

Percentage of BLASTx top hits with Gene Ontology (GO) mapping for ApeKI, PstI-MspI, and both combined, and the top two scoring biological processes and molecular functions.

Species or clade[i]	% hits with GO			Top two scoring biological processes[v]	Top two scoring molecular functions[v]
	A	PM	APM		
oc[ii]		100		memory, synaptic transmission	calcium channel activity, cation channel activity
bi[ii]		50.0		none scored	DNA binding
ca[ii]		70.0		neuropeptide signaling pathway, cellular metabolic process	zinc ion binding, sequence-specific DNA binding transcription factor activity
oc-bi-ca	100	75.0	83.3	metabolic process, cellular process	binding, nucleic acid binding
re[iii]	80.8			metabolic process, primary metabolic process	binding, catalytic activity
oc-bi-ca-re[iii]	75.0			transmembrane transport, ATP catabolic process	ligand-dependent nuclear receptor binding, ATP binding
fu	58.3	50.0	55.6	signal transduction, metabolic process	binding, transferase activity
oc-bi-ca-re-fu[iii]	64.4			cellular process, metabolic process	binding, catalytic activity
oc-bi-ca-fu[iv]		83.3		cellular metabolic process, metabolic process	binding, steroid dehydrogenase activity acting on the CH-OH group of donors, NAD, or NADP as acceptor
pi	62.9	65.9	64.0	cellular process, metabolic process	binding, nucleic acid binding
oc-bi-ca-re-fu-pi[iii]	84.0			cellular process, metabolic process	binding, nucleic acid binding
oc-bi-ca-fu-pi[iv]		80.3		cellular process, DNA metabolic process	binding, nucleic acid binding
ro	84.3	81.6	83.5	cellular process, metabolic process	binding, nucleic acid binding
co	83.8	84.0	83.8	cellular process, metabolic process	binding, nucleic acid binding
All specimens	38.9	26.3	32.5	cellular process, metabolic process	binding, nucleic acid binding

[i]Species are abbreviated using the first two letters: bi = *C. biennis*, ca = *C. carnana*, oc = *C. occidentalis*, re = *C. retiniana*, fu = *C. fumiferana*, pi = *C. pinus*, ro = *C. rosaceana*, co = *C. conflictana*.

[ii]For PstI-MspI (PM) only, because there were no BLASTx hits for ApeKI.

[iii]For ApeKI (A) only, because there were no *C. retiniana* specimens sequenced with PstI-MspI.

[iv]For PstI-MspI only, because the equivalent clade in ApeKI includes *C. retiniana.*

[v]Scores for GO terms are based on the number of sequences and their proximity to the node (term) being scored (Conesa et al., 2005).

Table 3-3

BLASTx top hits for *C. pinus* autapomorphic SNP sequences, including the querying sequence name (contig#|SNP position|alleles), and length (bp) of the query, and characteristics of the BLASTx top hit sequences.

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
Detoxification and immune response							
31706|1516|G_A	400	ENN79939	aael004336- partial	*Dendroctonus ponderosae*	7.28E-24	59	138
32261|8392|A_G	400	EHJ64029	alkaline nuclease precursor	*Danaus plexippus*	9.03E-37	87	85
79305|408|G_A	400	EHJ69795	atp-binding cassette sub-family g member 1-like	*Danaus plexippus*	3.95E-11	80	52
316254|166|G_A	366	EHJ71835	breast cancer metastasis-suppressor 1-like protein	*Danaus plexippus*	2.54E-15	89	48
15051|5261|A_G	400	EHJ69594	carboxylesterase	*Danaus plexippus*	9.18E-09	58	65
12163|705|C_G	400	ACZ97416	cytochrome p450	*Zygaena filipendulae*	2.36E-38	68	133
12163|706|T_C	400	ACZ97416	cytochrome p450	*Zygaena filipendulae*	2.36E-38	68	133
12163|719|A_G	400	ACZ97416	cytochrome p450	*Zygaena filipendulae*	2.25E-38	68	132
5448|6276|G_A	400	BAM20731	c-type partial	*Papilio polytes*	2.17E-58	90	103
26363|2675|G_A	400	EHJ77841	dipeptidase 1-like	*Danaus plexippus*	1.17E-53	93	118
26363|3643|G_A	400	EHJ77841	dipeptidase 1-like	*Danaus plexippus*	8.68E-56	99	133
75008|1134|C_T	400	AFN71166	glucose dehydrogenase	*Bombyx mori*	3.54E-43	74	129
62102|2431|G_C	374	ABW16859	prophenoloxidase 1	*Choristoneura fumiferana*	4.80E-44	87	91
18057|2535|T_G	400	ABX39545	single domain major allergen protein	*Helicoverpa armigera*	5.21E-26	86	69
236918|308|G_C	400	EHJ63682	sulfide:quinone mitochondrial-like	*Danaus plexippus*	1.13E-07	53	80
3819|3170|C_T	400	EHJ70493	e3 ubiquitin-protein ligase hyd-like	*Danaus plexippus*	9.39E-27	70	91
3819|3153|C_T	400	EHJ70493	e3 ubiquitin-protein ligase hyd-like	*Danaus plexippus*	1.84E-31	72	97
28210|3379|G_A	394	EHJ75811	e3 ubiquitin-protein ligase rififylin	*Danaus plexippus*	3.27E-06	76	59

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
23722|3646|T_C	400	EHJ65127	e3 ubiquitin-protein ligase rnf13-like	*Danaus plexippus*	3.06E-13	74	54
28103|1161|G_T	400	EHJ75060	ubiquitin protein ligase	*Danaus plexippus*	2.84E-16	94	57
2141|1237|G_A	400	XP_003706573	ubiquitin thioesterase trabid-like	*Megachile rotundata*	3.18E-14	81	48
5854|531|G_C	400	EHJ65343	ubiquitin-conjugating enzyme	*Danaus plexippus*	5.93E-10	68	76
177832|269|G_A	400	EHJ67483	ves g 5 allergen precursor	*Danaus plexippus*	3.50E-08	71	56
39405|1165|T_G	400	XP_001975062	vitamin k-dependent protein c	*Drosophila erecta*	1.61E-06	72	50
39405|1179|T_A	400	XP_001975062	vitamin k-dependent protein c	*Drosophila erecta*	1.61E-06	72	50
30316|581|C_T	400	EHJ74788	wd-repeat protein	*Danaus plexippus*	1.78E-21	98	51
64921|1941|A_G	400	EHJ77036	activating transcription factor- isoform b	*Danaus plexippus*	1.11E-20	87	55
252513|348|G_T	400	EHJ71292	basigin	*Danaus plexippus*	1.31E-24	96	51
406|2120|T_A	400	EHJ73664	c-maf-inducing protein	*Danaus plexippus*	2.57E-21	98	50
13015|6161|C_T	400	XP_970465	leucine-rich repeats and immunoglobulin-like domains protein 2	*Tribolium castaneum*	2.78E-10	81	43
13015|6162|A_G	400	XP_970465	leucine-rich repeats and immunoglobulin-like domains protein 2	*Tribolium castaneum*	7.53E-11	81	44
1164|1280|C_T	400	EHJ67058	tyrosine-protein kinase csk	*Danaus plexippus*	1.83E-24	87	57
51479|1550|G_C	400	AFG30048	heat shock protein 90 beta	*Bombyx mori*	1.25E-34	87	86
129111|463|G_T	400	EHJ68095	hypothetical protein KGM_02923	*Danaus plexippus*	1.24E-18	90	54
Metabolism							
810|2209|G_T	400	EHJ69331	3-hydroxyacyl-dehydrogenase type-2	*Danaus plexippus*	9.48E-11	86	38
810|2148|T_C	400	EHJ69331	3-hydroxyacyl-dehydrogenase type-2	*Danaus plexippus*	1.94E-21	87	56
78046|1500|A_G	400	XP_003649139	abhydrolase domain containing 11	*Thielavia terrestris NRRL 8126*	1.21E-25	65	119
57779|1282|G_A	400	EHJ72683	guanylate cyclase 32e-like	*Danaus plexippus*	6.73E-25	80	81
11586|4744|C_T	400	EHJ68120	high density lipoprotien binding protein vigilin	*Danaus plexippus*	1.42E-17	89	58

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
17966|6415|C_T	400	EHJ69667	insulin-like growth factor 2 (igf2) mRNA binding protein	*Danaus plexippus*	8.34E-37	89	88
1038|224|G_A	400	EHJ74496	krueppel-like factor 15	*Danaus plexippus*	1.61E-37	94	69
35835|806|G_A	400	AAC62229	lipase 3-like	*Plodia interpunctella*	1.12E-46	74	133
334723|214|G_A	384	EHJ74292	low quality protein: wd repeat and fyve domain-containing protein 3-like	*Danaus plexippus*	1.39E-43	100	77
65513|458|C_T	400	EHJ74816	luciferase	*Danaus plexippus*	5.17E-21	77	68
5910|2495|A_C	400	EGI57828	pancreatic triacylglycerol lipase-like	*Acromyrmex echinatior*	1.90E-14	67	71
87740|255|G_A	400	EHJ71299	tafazzin homolog	*Danaus plexippus*	3.26E-22	79	69
23110|4179|C_T	400	NP_001036959	membrane metallo-endopeptidase-like 1-like	*Bombyx mori*	2.48E-44	84	101
Sensory Perception and Motor Control							
11514|4173|G_C	400	EHJ67823	nuclear receptor subfamily 2 group f member 6	*Danaus plexippus*	1.95E-16	94	38
314|7849|G_C	400	EHJ67200	shaggy	*Danaus plexippus*	4.59E-14	95	41
34822|926|A_G	400	AFA55158	beta adrenergic-like octopamine receptor partial	*Trichoplusia ni*	9.98E-54	95	100
44154|209|G_C	400	XP_002106959	potassium voltage-gated channel protein eag	*Drosophila simulans*	1.84E-36	100	65
42248|2507|C_T	400	EHJ69137	protein stoned-a	*Danaus plexippus*	2.60E-22	77	86
17662|3894|C_T	400	EHJ71750	protein turtle	*Danaus plexippus*	2.66E-33	95	68
13869|2164|C_T	400	EHJ65422	putative otopetrin	*Danaus plexippus*	4.43E-13	81	55
22025|987|T_G	400	EHJ65424	hypothetical protein KGM_05456	*Danaus plexippus*	6.83E-30	78	85
6938|9854|G_C	400	EHJ67191	semaphorin 2a	*Danaus plexippus*	2.72E-27	93	64
11865|4478|C_T	400	EHJ79117	ankyrin repeat and fyve domain-containing protein 1-like	*Danaus plexippus*	2.33E-27	94	79
21316|827|G_T	400	EHJ73805	ankyrin repeat domain-containing	*Danaus plexippus*	3.08E-33	94	74
26027|4713|G_A	400	EHJ74853	intraflagellar transport protein 140 homolog	*Danaus plexippus*	4.26E-45	93	89
13874|2670|G_A	400	AFC91751	odorant receptor partial	*Cydia pomonella*	9.78E-10	81	38

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
29003|1423|G_C	400	ACD85125	gustatory receptor 45	*Bombyx mori*	6.06E-08	49	136
3096|1104|C_T	400	BAM19881	serine threonine-protein kinase doa	*Papilio xuthus*	1.00E-19	92	50
275383|135|G_A	335	EHJ64604	low quality protein: otoferlin-like	*Danaus plexippus*	6.87E-19	74	70
Morphology							
32982|3937|T_C	400	EHJ69319	ankyrin repeat domain-containing protein 12	*Danaus plexippus*	9.12E-20	97	47
13123|4039|C_T	400	EHJ63549	hormone receptor-like in isoform d	*Danaus plexippus*	5.01E-39	87	90
57267|1517|G_A	400	EHJ77642	glycylpeptide n-tetradecanoyltransferase 2	*Danaus plexippus*	1.12E-87	96	133
18632|4220|G_A	400	EHJ74076	nonclathrin coat protein gamma1-cop	*Danaus plexippus*	8.43E-43	98	96
66780|870|C_T	400	XP_001870024	xylosyltransferase oxt	*Culex quinquefasciatus*	2.98E-11	79	49
35452|1052|C_T	400	BAM19801	oligosaccharyl transferase	*Papilio xuthus*	3.13E-35	96	64
13906|10419|A_C	400	EHJ74017	scabrous protein	*Danaus plexippus*	5.26E-13	82	57
7181|2075|G_A	400	EHJ68177	trithorax group protein osa	*Danaus plexippus*	1.76E-44	94	133
31700|2177|C_G	400	XP_002423712	trithorax group protein osa	*Pediculus humanus corporis*	1.42E-29	100	58
Cell cycle							
129353|607|T_G	400	EHJ67522	cd9 antigen	*Danaus plexippus*	3.99E-34	95	63
18680|569|A_G	400	BAM18612	cdc2-related kinase	*Papilio xuthus*	4.23E-43	98	77
3438|500|G_A	400	EHJ76144	cell division cycle protein 23 homolog	*Danaus plexippus*	1.10E-24	90	63
39100|2026|C_G	304	BAM19369	cyclin g	*Papilio xuthus*	4.58E-11	81	43
115912|1085|A_G	305	EHJ68425	sin3a-associated protein sap130	*Danaus plexippus*	4.42E-19	95	46
37011|4128|T_G	400	BAM20371	spliceosomal protein sap	*Papilio polytes*	1.11E-20	96	50
29348|2969|C_T	400	EHJ68113	nipped-b-like protein	*Danaus plexippus*	1.33E-30	100	64
16950|9758|C_A	400	EHJ67381	hypothetical protein KGM_13834	*Danaus plexippus*	2.65E-12	59	129

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
16950|13361|C_T	400	EHJ67381	hypothetical protein KGM_13834	*Danaus plexippus*	2.66E-50	84	134
97200|650|C_T	400	EHJ63744	hypothetical protein KGM_18067	*Danaus plexippus*	2.89E-12	97	36
20419|8581|C_T	400	EHJ68087	insulin-like growth factor-binding protein complex acid labile chain	*Danaus plexippus*	4.84E-56	94	128
8839|433|C_T	400	XP_394148	proto-oncogene tyrosine-protein kinase ros	*Apis mellifera*	1.92E-12	49	128
152698|551|A_C	400	EHJ63897	huntingtin interacting protein 1	*Danaus plexippus*	8.40E-09	81	43
13687|6086|G_A	400	EHJ76256	hypothetical protein KGM_00312	*Danaus plexippus*	1.19E-50	85	118
68678|472|A_G	400	NP_001036936	cell cycle checkpoint kinase 2	*Bombyx mori*	6.32E-29	96	58
6991|4845|T_A	400	EHJ70533	myotubularin-related protein 9-like	*Danaus plexippus*	8.12E-31	86	74
29263|4322|G_A	400	EHJ64633	ww domain-containing oxidoreductase	*Danaus plexippus*	1.11E-30	70	96
Cell movement							
4668|1127|G_A	400	EHJ77113	axonemal dynein heavy chain	*Danaus plexippus*	7.24E-43	93	83
48181|2471|C_T	400	XP_971508	probable e3 ubiquitin-protein ligase mycbp2	*Tribolium castaneum*	1.36E-24	67	71
12528|6677|C_T	400	EHJ77583	protein ect2-like isoform 3	*Danaus plexippus*	1.98E-14	83	54
Intracellular and cell-cell communication							
1414|2460|T_A	400	EHJ63441	agrin-like isoform 1	*Danaus plexippus*	3.90E-57	91	118
105121|503|A_T	400	XP_001851471	ctl-like protein 1-like	*Culex quinquefasciatus*	8.38E-10	69	46
14961|1448|T_A	400	EHJ65570	glutamate receptor 1-like	*Danaus plexippus*	1.39E-18	93	47
43970|2493|G_A	400	EHJ64934	hypothetical protein KGM_00665	*Danaus plexippus*	3.34E-33	66	133
16139|3177|C_A	400	ENN78849	inactive ubiquitin carboxyl-terminal hydrolase 54	*Dendroctonus ponderosae*	2.08E-08	56	128
36319|278|C_A	400	EHJ67267	ionotropic receptor isoform f	*Danaus plexippus*	3.15E-18	83	53
77598|995|C_T	400	EHJ69501	isoform p	*Danaus plexippus*	3.24E-40	100	70
1699|246|T_C	400	EHJ68109	isoform v	*Danaus plexippus*	2.48E-10	77	48

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
1707|20105|G_A	400	EHJ68109	isoform x	*Danaus plexippus*	7.66E-51	83	132
21178|6358|A_G	400	EHJ64475	laminin subunit gamma-3	*Danaus plexippus*	1.71E-45	94	84
158990|449|T_C	400	NP_001127749	neuropeptide receptor a33	*Bombyx mori*	1.98E-11	80	47
158990|453|A_T	400	NP_001127749	neuropeptide receptor a33	*Bombyx mori*	7.72E-10	80	47
1414|2344|C_A	400	EHJ63441	putative agrin	*Danaus plexippus*	1.21E-47	92	102
3447|4720|C_T	400	XP_003700653	rab3 gtpase-activating protein catalytic subunit	*Megachile rotundata*	2.44E-14	68	86
8285|2758|C_T	400	EHJ69272	solute carrier family 12 member 6	*Danaus plexippus*	5.11E-22	98	50
22351|3849|T_A	400	EHJ63446	follistatin-related protein 1- partial	*Danaus plexippus*	3.34E-33	84	77
5606|2555|G_A	400	AEW46914	seminal fluid protein cssfp066	*Chilo suppressalis*	1.07E-16	91	48
54306|2319|C_T	400	EHJ70995	steroid receptor-interacting snf2 domain protein	*Danaus plexippus*	2.06E-25	96	65
233742|410|G_A	400	O76202	ultraspiracle protein	*Choristoneura fumiferana*	2.26E-23	81	71
233742|416|C_T	400	O76202	ultraspiracle protein	*Choristoneura fumiferana*	3.37E-23	81	71
15055|1375|G_A	400	EHJ68098	adenylate cyclase	*Danaus plexippus*	3.15E-13	78	52
159198|320|G_A	400	EHJ68158	arf-gap with dual ph domain-containing protein 1-like	*Danaus plexippus*	1.10E-28	90	63
7316|6684|G_T	400	EHJ76073	fused1 protein	*Danaus plexippus*	1.52E-41	96	52
18805|2797|C_T	400	EHJ76804	hypothetical protein KGM_01006	*Danaus plexippus*	1.25E-07	67	62
18805|1113|C_T	400	EHJ76804	hypothetical protein KGM_01006	*Danaus plexippus*	1.73E-12	74	59
7062|6554|G_A	400	EHJ66187	hypothetical protein KGM_06010	*Danaus plexippus*	1.07E-11	83	61
42713|5808|C_T	400	EHJ77737	hypothetical protein KGM_07530	*Danaus plexippus*	2.61E-26	94	55
5048|248|C_A	400	EHJ65790	hypothetical protein KGM_07941	*Danaus plexippus*	2.37E-06	65	60
10917|7064|T_C	400	EHJ77981	hypothetical protein KGM_17378	*Danaus plexippus*	5.85E-19	89	48
13281|925|C_T	400	EHJ70388	importin beta-3	*Danaus plexippus*	1.48E-30	87	73
39849|2970|T_A	400	EHJ77073	isoform h	*Danaus plexippus*	4.63E-46	90	99
24204|7585|C_T	400	EHJ77209	low-density lipoprotein receptor	*Danaus plexippus*	2.87E-20	94	53

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
72988|1170|C_G	400	CAY54150	monocarboxylate transporter-like protein	*Heliconius melpomene*	7.82E-31	95	64
31368|1573|C_T	400	BAM18531	protein lin-7 homolog b-like	*Papilio xuthus*	3.16E-38	100	70
5142|7646|C_T	400	XP_002428547	protein patched	*Pediculus humanus corporis*	2.13E-15	76	82
33886|104|G_C	304	XP_968541	protein vac14 homolog	*Tribolium castaneum*	3.30E-10	76	42
252783|310|C_G	400	EHJ67824	proton-coupled folate transporter-like	*Danaus plexippus*	4.76E-25	95	60
22360|912|G_A	400	EHJ67834	scy1-like protein 2-like	*Danaus plexippus*	6.07E-31	88	101
53501|787|A_G	400	NP_001233204	soluble guanylyl cyclase alpha-1 subunit	*Bombyx mori*	4.57E-22	84	65
95016|612|C_G	400	EHJ72348	sugar transporter erd6-like 6	*Danaus plexippus*	3.14E-30	79	91
68696|158|G_A	358	EHJ66669	coatomer subunit beta	*Danaus plexippus*	2.81E-18	96	57
68696|206|C_T	400	EHJ66669	coatomer subunit beta	*Danaus plexippus*	4.36E-18	96	57
53326|2798|C_T	400	BAM18615	golgi reassembly stacking protein 2	*Papilio xuthus*	4.18E-16	67	82
53326|2831|A_G	400	BAM18615	golgi reassembly stacking protein 2	*Papilio xuthus*	4.12E-15	68	92
6633|3484|A_G	400	EHJ74239	golgi to er traffic protein 4 homolog	*Danaus plexippus*	8.05E-44	87	95
166911|745|G_A	322	EHJ67030	rab6 gtpase activating gapcena	*Danaus plexippus*	1.59E-23	90	63
279985|238|C_G	400	EHJ69304	signal transducing adapter molecule 1-like	*Danaus plexippus*	4.77E-10	94	35
279985|254|G_A	400	EHJ69304	signal transducing adapter molecule 1-like	*Danaus plexippus*	6.20E-10	94	35
Gene expression							
12844|3622|A_G	400	EHJ64547	additional sex combs (asx)	*Danaus plexippus*	1.49E-27	88	134
102462|872|A_C	400	EHJ77636	arylhydrocarbon receptor nuclear translocator-like protein b	*Danaus plexippus*	6.69E-11	80	45
93200|600|T_C	400	EHJ68194	capicua protein	*Danaus plexippus*	5.91E-06	64	68
42290|1818|C_G	400	EHJ68697	eukaryotic translation initiation factor 4 2	*Danaus plexippus*	8.96E-27	72	117
45230|1242|C_G	400	EHJ66491	gata transcription factor gatac	*Danaus plexippus*	4.98E-27	73	126

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
45230|1222|T_C	400	EHJ66491	gata transcription factor gatac	*Danaus plexippus*	4.85E-31	74	133
45230|352|G_A	400	EGI57433	gata-binding factor a	*Acromyrmex echinatior*	1.49E-05	62	66
103225|377|A_G	400	XP_002058096	GJ15893	*Drosophila virilis*	5.98E-05	66	45
9315|15431|C_G	400	EHJ64926	RNA polymerase-associated protein ctr9 homolog	*Danaus plexippus*	2.79E-23	86	59
7198|957|G_A	400	EHJ66129	xpg-like endonuclease	*Danaus plexippus*	6.64E-23	76	60
7198|999|C_A	400	EHJ66129	xpg-like endonuclease	*Danaus plexippus*	6.63E-23	76	60
398223|197|C_T	296	EHJ77321	yeats domain-containing protein 2	*Danaus plexippus*	2.67E-11	69	56
30004|232|A_G	400	EHJ67957	zinc finger and btb domain-containing protein 3	*Danaus plexippus*	1.31E-21	84	64
1269|17633|G_A	400	EHJ73037	zinc finger homeobox protein 3	*Danaus plexippus*	1.86E-49	96	117
5874|14941|G_A	400	ELK10737	zinc finger protein 75d-like	*Pteropus alecto*	2.16E-13	61	68
17128|7340|C_G	400	EHJ63786	cg12701 cg12701-pa	*Danaus plexippus*	1.09E-70	94	134
51134|1324|G_C	400	EHJ63883	forkhead box protein k2-like	*Danaus plexippus*	4.30E-26	92	57
131228|636|C_T	400	EHJ67685	gag-like protein	*Danaus plexippus*	1.06E-37	69	133
146015|119|C_T	319	YP_003517871	gag-like protein	*Lymantria xylina MNPV*	1.95E-25	72	85
36128|2739|G_A	400	BAM17991	cyclophilin a	*Papilio xuthus*	1.81E-77	99	117
2748|4155|G_A	400	NP_001040301	glycosyl-phosphatidyl-inositol-anchored protein	*Bombyx mori*	3.89E-20	97	46
93559|1381|A_T	291	EFA09485	histone-lysine n-methyltransferase setmar	*Tribolium castaneum*	4.61E-23	67	87
13046|1606|A_T	400	NP_001093282	histone-lysine n-methyltransferase setmar	*Bombyx mori*	6.78E-50	75	133
436135|110|C_T	243	EHJ78150	isoform a	*Danaus plexippus*	4.54E-17	87	54
1189|7848|C_A	400	EHJ77349	isoform a	*Danaus plexippus*	4.32E-54	91	134
6268|1349|C_T	400	EHJ67343	isoform a	*Danaus plexippus*	6.68E-27	94	57
43886|1199|G_A	400	EHJ65955	isoform c	*Danaus plexippus*	3.97E-14	100	38
5909|2625|G_T	400	EHJ63326	menin-like	*Danaus plexippus*	4.10E-73	93	132
5909|2602|C_T	400	EHJ63326	low quality protein: menin-like	*Danaus plexippus*	6.77E-75	93	133

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
72713|1047|G_T	400	ACU24710	maternal protein pumilio	*Bombyx mori*	4.04E-29	98	58
32329|1701|C_G	400	EHJ67256	mediator of RNA polymerase ii transcription subunit 24-like	*Danaus plexippus*	1.96E-38	76	97
13506|564|A_T	400	EHJ71132	mitochondrial ribosomal protein l53	*Danaus plexippus*	5.84E-17	93	43
2924|254|C_T	400	EHJ71125	nucleoside diphosphate kinase 7 isoform 2	*Danaus plexippus*	4.79E-27	88	79
38395|1309|A_G	400	EHJ64537	prip interacting pimt	*Danaus plexippus*	1.29E-11	83	43
78022|984|A_G	400	EFA13284	protein	*Tribolium castaneum*	7.40E-22	62	103
17376|767|A_T	400	EFR21509	protein scai-like	*Anopheles darlingi*	5.08E-21	72	76
318790|293|G_C	333	AGC92703	protein shuttle craft	*Heliconius erato*	9.18E-15	62	80
37003|803|G_C	400	EHJ73473	sumo ligase	*Danaus plexippus*	1.22E-10	81	60
30004|250|G_C	400	EHJ67957	transcription factor rsv1	*Danaus plexippus*	1.00E-18	85	57
4878|923|G_A	400	NP_001040313	transcription factor-like protein	*Bombyx mori*	1.28E-15	87	78
158470|446|T_C	400	EHJ75707	u1 small nuclear ribonucleoprotein c	*Danaus plexippus*	8.37E-35	96	64
158470|464|G_A	400	EHJ75707	u1 small nuclear ribonucleoprotein c	*Danaus plexippus*	8.36E-29	96	57
65465|298|C_T	400	EHJ64026	integrator complex subunit 4-like	*Danaus plexippus*	4.69E-20	83	73
54558|456|A_C	400	EMR08568	peptidyl-tRNA hydrolase mitochondrial	*Pneumocystis murina B123*	1.94E-21	58	117
9823|9076|A_T	400	EHJ70532	protein kinase c-binding protein 1	*Danaus plexippus*	5.45E-19	87	41
76475|1706|A_C	319	EHJ67667	cg3106 cg3106-pa	*Danaus plexippus*	2.95E-24	83	65
88248|1324|A_T	400	EHJ63874	putative ribonucleoprotein	*Danaus plexippus*	6.48E-22	92	53
88248|1330|A_C	400	EHJ63874	putative ribonucleoprotein	*Danaus plexippus*	1.92E-23	92	55
25459|2195|C_A	400	EHJ70079	hypothetical protein KGM_02189	*Danaus plexippus*	7.08E-15	84	51
140062|886|G_A	333	EHJ71248	hypothetical protein KGM_08698	*Danaus plexippus*	4.21E-08	70	48
316|4909|C_T	400	EHJ67195	hypothetical protein KGM_10806	*Danaus plexippus*	1.81E-18	81	132
2116|8165|C_T	400	EHJ70635	hypothetical protein KGM_15031	*Danaus plexippus*	1.65E-22	62	127

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
2116|7916|C_G	400	EHJ70635	hypothetical protein KGM_15031	*Danaus plexippus*	4.43E-18	72	100
259012|324|C_G	400	EHJ79243	hypothetical protein KGM_15411	*Danaus plexippus*	1.20E-08	54	101
Mobile genetic elements							
23627|900|A_G	369	AAQ57129	endonuclease and reverse transcriptase-like protein	*Bombyx mori*	5.49E-22	62	106
11693|3141|G_A	400	XP_003241651	endonuclease-reverse transcriptase	*Acyrthosiphon pisum*	4.30E-17	76	76
20678|3219|G_A	282	ADI61811	endonuclease-reverse transcriptase	*Bombyx mori*	3.25E-20	87	55
25765|2050|C_T	400	EHJ74581	endonuclease-reverse transcriptase	*Danaus plexippus*	2.88E-07	66	62
311998|100|A_T	300	ADI61811	endonuclease-reverse transcriptase	*Bombyx mori*	8.70E-30	89	59
3774|2137|T_C	400	EHJ63466	endonuclease-reverse transcriptase	*Danaus plexippus*	4.83E-13	78	57
43490|2456|A_G	400	EHJ73538	endonuclease-reverse transcriptase	*Danaus plexippus*	1.22E-13	75	64
43490|2458|A_G	400	EHJ73538	endonuclease-reverse transcriptase	*Danaus plexippus*	1.22E-13	75	64
43490|2480|C_A	400	EHJ73538	endonuclease-reverse transcriptase	*Danaus plexippus*	1.22E-13	75	64
444747|91|A_G	231	EHJ77256	endonuclease-reverse transcriptase	*Danaus plexippus*	3.15E-20	83	60
4567|2808|G_A	400	ADI61811	endonuclease-reverse transcriptase	*Bombyx mori*	4.40E-13	78	42
49725|979|G_A	400	EHJ74922	endonuclease-reverse transcriptase -e01	*Danaus plexippus*	4.44E-04	63	44
52854|302|C_T	400	ABO45239	endonuclease-reverse transcriptase	*Ostrinia nubilalis*	9.78E-14	92	40
52854|337|T_C	400	ADI61819	endonuclease-reverse transcriptase	*Bombyx mori*	3.23E-13	81	49
52854|345|T_C	400	ADI61819	endonuclease-reverse transcriptase	*Bombyx mori*	3.32E-13	81	49
62233|924|G_A	400	ADI61824	endonuclease-reverse transcriptase	*Bombyx mori*	1.12E-19	64	48
62254|378|G_A	400	EHJ65125	endonuclease-reverse transcriptase	*Danaus plexippus*	1.33E-37	74	103
83990|448|T_A	400	EHJ73538	endonuclease-reverse transcriptase	*Danaus plexippus*	7.21E-30	84	82
8403|1846|G_C	400	EHJ73538	endonuclease-reverse transcriptase	*Danaus plexippus*	7.03E-05	56	55
8529|425|G_A	400	ADI61810	endonuclease-reverse transcriptase	*Bombyx mori*	6.52E-42	82	113

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
78030|1387|C_T	400	EFA01934	retrotransposon-like family member (retr-1)-like	*Tribolium castaneum*	9.61E-08	54	108
82674|762|C_T	400	BAC82595	reverse transcriptase	*Anopheles gambiae*	7.51E-07	57	80
48998|2151|A_C	400	EHJ73416	reverse transcriptase	*Danaus plexippus*	1.15E-09	62	83
48998|2107|A_T	400	EHJ73416	reverse transcriptase	*Danaus plexippus*	7.43E-08	63	68
5217|2887|G_C	400	BAD86652	reverse transcriptase	*Bombyx mori*	4.59E-11	79	44
117016|542|T_C	400	EFA13284	RNA-directed DNA polymerase from mobile element jockey-like	*Tribolium castaneum*	4.53E-14	52	88
17222|1347|C_G	400	S08405	S08405hypothetical protein 2 - silkworm transposon mag	*Bombyx mori*	2.01E-21	69	108
93559|1367|G_A	305	EFA09485	transposase	*Tribolium castaneum*	3.85E-23	67	87
57121|2398|C_G	400	EHJ70885	pol-like protein	*Danaus plexippus*	1.26E-09	54	110
92511|419|C_T	400	BAB21511	TRAS3	*Bombyx mori*	2.07E-17	60	125
92511|427|C_T	400	BAB21511	TRAS3	*Bombyx mori*	1.18E-18	60	128
13749|3466|G_A	400	EFA13284	hypothetical protein TcasGA2_TC010304	*Tribolium castaneum*	8.42E-04	73	38
482582|48|A_T	184	EHJ64305	putative toll	*Danaus plexippus*	9.57E-20	85	61
Multiple functions							
19519|3648|G_A	400	EHJ69771	hypothetical protein KGM_06966	*Danaus plexippus*	3.32E-16	94	51
23862|5542|C_T	400	EHJ69309	hypothetical protein KGM_10889	*Danaus plexippus*	4.55E-07	66	36
41643|3317|A_T	400	EHJ66848	hypothetical protein KGM_12071	*Danaus plexippus*	1.21E-16	82	84
5277|5940|A_T	400	EHJ70940	hypothetical protein KGM_14803	*Danaus plexippus*	1.14E-21	61	139
1915|2201|G_A	400	EHJ69186	hypothetical protein KGM_18565	*Danaus plexippus*	3.38E-14	61	89
19628|3718|G_A	400	EHJ74066	hypothetical protein KGM_18585	*Danaus plexippus*	1.65E-17	88	80
103195|228|G_A	400	EHJ75148	hypothetical protein KGM_21424	*Danaus plexippus*	8.11E-06	70	47
103195|229|T_G	400	EHJ75148	hypothetical protein KGM_21424	*Danaus plexippus*	8.11E-06	70	47

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
103195|230|C_A	400	EHJ75148	hypothetical protein KGM_21424	*Danaus plexippus*	8.01E-06	70	47
Unknown							
336532|187|C_T	387	EHJ68065	hypothetical protein KGM_01227	*Danaus plexippus*	2.53E-10	76	96
24861|2911|C_T	400	EHJ76775	hypothetical protein KGM_02205	*Danaus plexippus*	9.75E-11	94	35
33387|1939|A_G	400	EHJ70759	hypothetical protein KGM_03388	*Danaus plexippus*	7.35E-18	86	69
16102|8161|G_A	400	EHJ75393	hypothetical protein KGM_03603	*Danaus plexippus*	4.05E-09	51	137
1118|2935|C_A	400	EHJ77612	hypothetical protein KGM_04635	*Danaus plexippus*	3.54E-09	71	53
1118|2941|C_A	400	EHJ77612	hypothetical protein KGM_04635	*Danaus plexippus*	1.47E-09	71	53
300731|44|T_A	244	EHJ67908	hypothetical protein KGM_05347	*Danaus plexippus*	1.29E-16	65	79
66491|866|T_C	400	EHJ71911	hypothetical protein KGM_06038	*Danaus plexippus*	6.75E-08	64	62
101784|657|C_T	400	EHJ66820	hypothetical protein KGM_06825	*Danaus plexippus*	9.57E-04	72	36
36203|2086|G_A	400	EHJ77736	hypothetical protein KGM_07539	*Danaus plexippus*	3.10E-38	83	86
3175|6557|A_G	400	EHJ66674	hypothetical protein KGM_08774	*Danaus plexippus*	1.52E-22	86	61
36318|816|C_T	400	EHJ67267	hypothetical protein KGM_09992	*Danaus plexippus*	4.77E-18	82	52
5175|13025|G_T	400	EHJ69351	hypothetical protein KGM_10862	*Danaus plexippus*	1.87E-31	71	139
72664|1110|G_A	400	EHJ70366	hypothetical protein KGM_13647	*Danaus plexippus*	2.59E-07	95	47
11520|1717|A_G	400	EHJ78446	hypothetical protein KGM_16284	*Danaus plexippus*	9.06E-22	66	133
23215|4138|G_A	400	EHJ65697	hypothetical protein KGM_16360	*Danaus plexippus*	4.63E-44	94	88
61607|2271|T_G	400	EHJ63217	hypothetical protein KGM_16446	*Danaus plexippus*	5.87E-08	75	49
8478|4008|G_A	400	EHJ77994	hypothetical protein KGM_17386	*Danaus plexippus*	5.92E-14	77	59
39804|843|C_T	400	EHJ74077	hypothetical protein KGM_18644	*Danaus plexippus*	3.13E-37	94	86
15987|2212|G_A	400	EHJ77895	hypothetical protein KGM_18674	*Danaus plexippus*	3.92E-20	80	82
20380|12280|T_C	400	EHJ63611	hypothetical protein KGM_19900	*Danaus plexippus*	6.97E-05	63	58

Sequence name	Sequence length (bp)	Hit accession number	Hit sequence description	Species	E-Value	Identity (%)	Alignment length (aa)
20380|9702|G_A	400	EHJ63612	hypothetical protein KGM_19909	*Danaus plexippus*	4.89E-35	68	135
27747|1835|C_T	400	EHJ68869	hypothetical protein KGM_19966	*Danaus plexippus*	7.33E-10	97	89
27747|1874|C_T	400	EHJ68869	hypothetical protein KGM_19966	*Danaus plexippus*	5.69E-12	98	109
15105|1182|G_A	400	EHJ78299	hypothetical protein KGM_22705	*Danaus plexippus*	3.22E-26	79	133
31317|1650|C_A	400	EFA09086	hypothetical protein TcasGA2_TC006806	*Tribolium castaneum*	4.54E-04	66	39
31317|1656|C_A	400	EFA09086	hypothetical protein TcasGA2_TC006806	*Tribolium castaneum*	4.36E-04	66	39
14824|2990|G_A	400	XP_003691339	uncharacterized protein LOC100871724	*Apis florea*	2.58E-14	67	59
23862|4440|C_A	400	XP_003738758	uncharacterized protein LOC100903590	*Metaseiulus occidentalis*	4.77E-04	64	53
2059|6116|G_A	400	XP_003740256	uncharacterized protein LOC100905519	*Metaseiulus occidentalis*	2.38E-04	49	67
101011|165|T_C	365	XP_003740256	uncharacterized protein LOC100905519	*Metaseiulus occidentalis*	3.88E-19	56	121
169141|497|T_A	400	XP_003740256	uncharacterized protein LOC100905519	*Metaseiulus occidentalis*	3.01E-12	56	104
51949|3463|T_C	400	XP_003740256	uncharacterized protein LOC100905519	*Metaseiulus occidentalis*	1.32E-20	59	124
1577|65|G_C	265	ABR17332	unknown	*Picea sitchensis*	1.25E-10	67	56
13419|5303|G_T	400	EHJ66659	coiled-coil domain-containing protein agap005037	*Danaus plexippus*	7.14E-07	85	35
13421|3427|T_C	400	EHJ66659	coiled-coil domain-containing protein agap005037	*Danaus plexippus*	1.06E-30	80	85
13421|3429|T_C	400	EHJ66659	coiled-coil domain-containing protein agap005037	*Danaus plexippus*	1.02E-30	80	85
13421|3430|C_A	400	EHJ66659	coiled-coil domain-containing protein agap005037	*Danaus plexippus*	2.44E-31	80	86
12483|4747|G_A	400	EHJ64428	flj12716-like protein	*Danaus plexippus*	6.88E-31	72	133
7137|2175|G_A	400	EHJ77035	unknown	*Danaus plexippus*	2.19E-24	98	50
25054|5071|C_A	400	BAM18578	similar to CG30108	*Papilio xuthus*	1.30E-08	80	42
60277|2368|G_A	400	XP_003695401	upf0505 protein c16orf62 homolog	*Apis florea*	4.66E-09	77	40

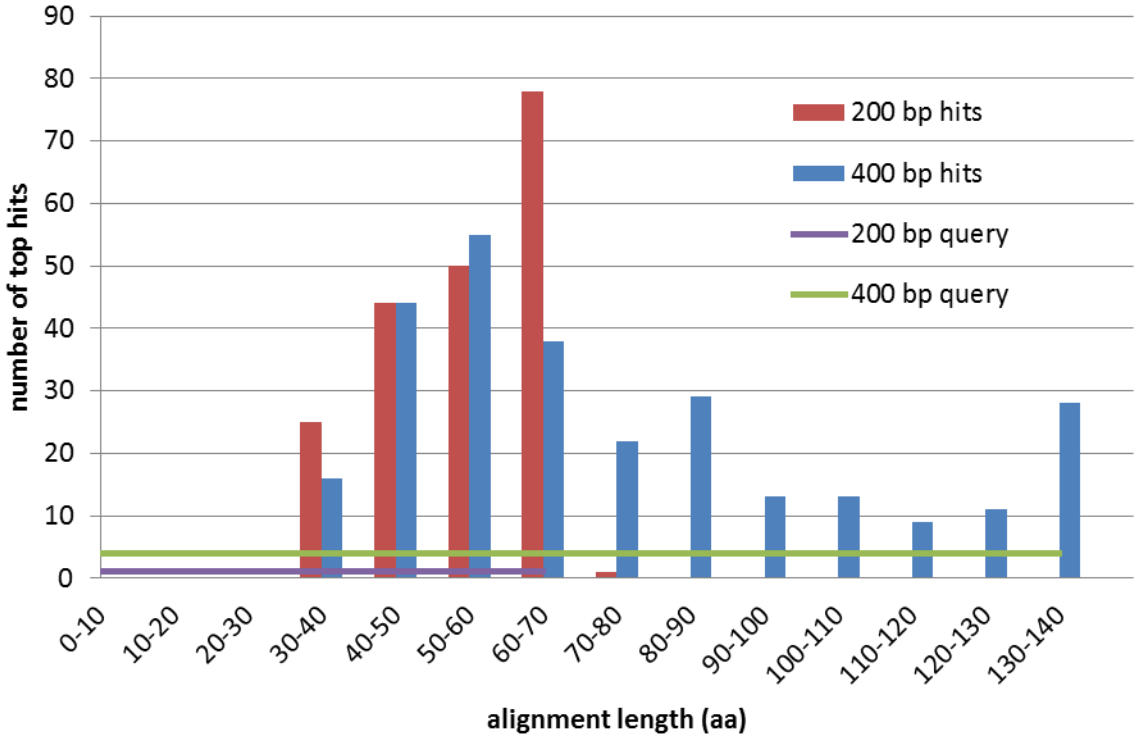


Figure 3-1. Distribution of alignment lengths of BLASTx top hits to 200 bp and 400 bp *C. pinus* autapomorphic SNP sequences. Autapomorphic SNPs from the ApeKI and PstI-MspI datasets were combined. The maximum lengths of the querying sequence are shown along the x-axis (green and purple lines). Top hits beyond the maximum alignment length are from gaps in the alignment.

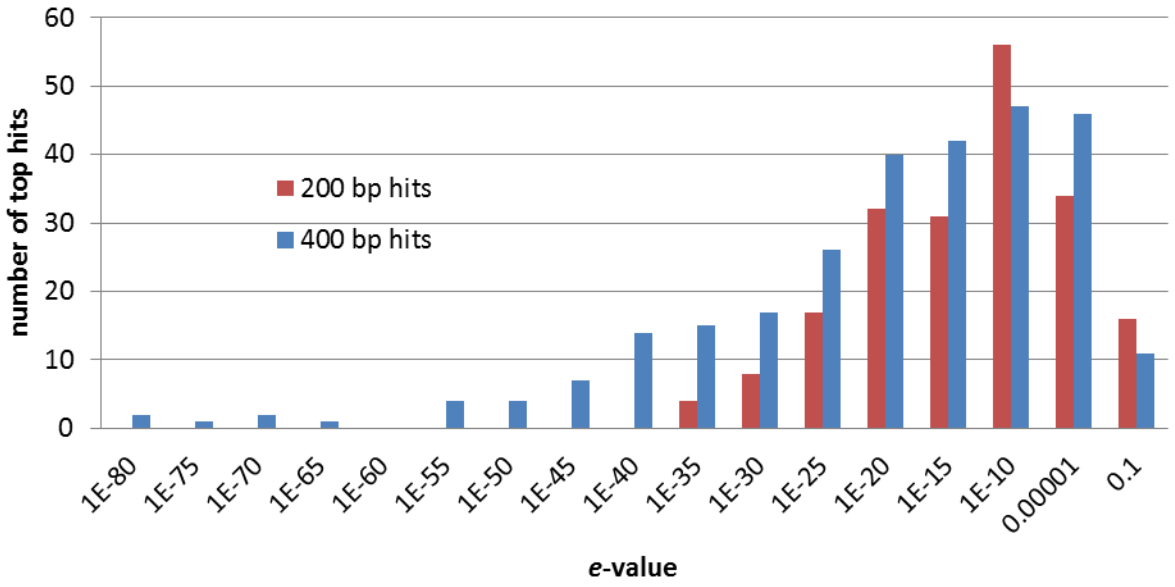


Figure 3-2. Distribution of *e*-values of BLASTx top hits for 200 bp (mean 1.58E-05) and 400 bp (mean 1.48E-05) *C. pinus* SNP sequences.

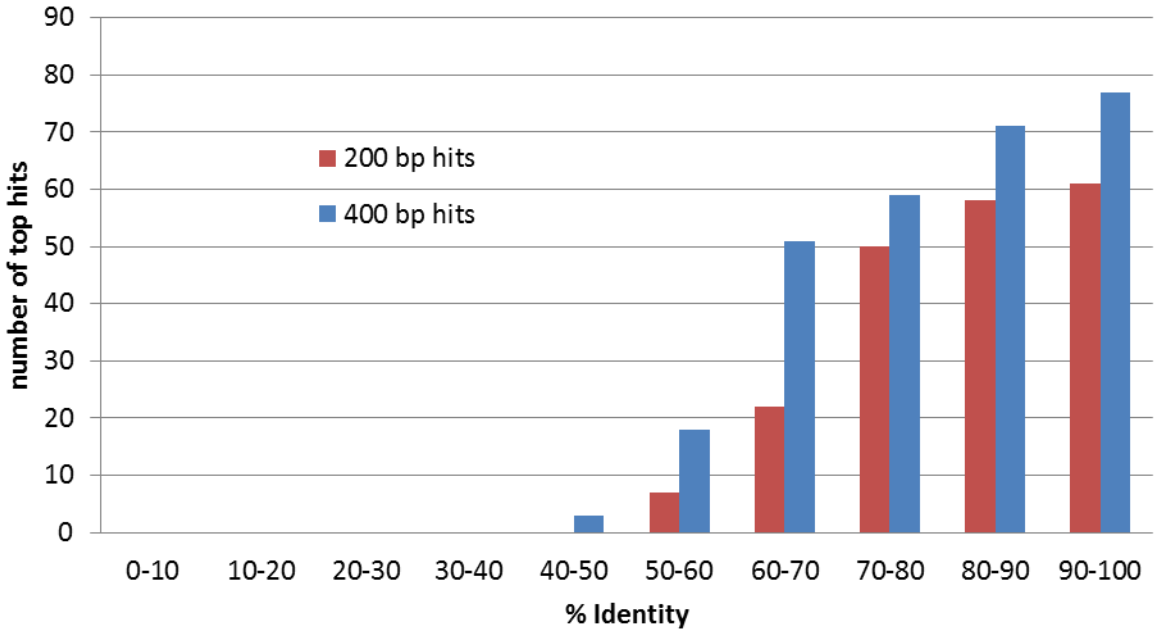


Figure 3-3. Distribution of percent identity of BLASTx top hit alignment lengths (not query lengths) for 200 bp (mean 83.11%) and 400 bp (mean 80.14%) *C. pinus* SNP sequences.

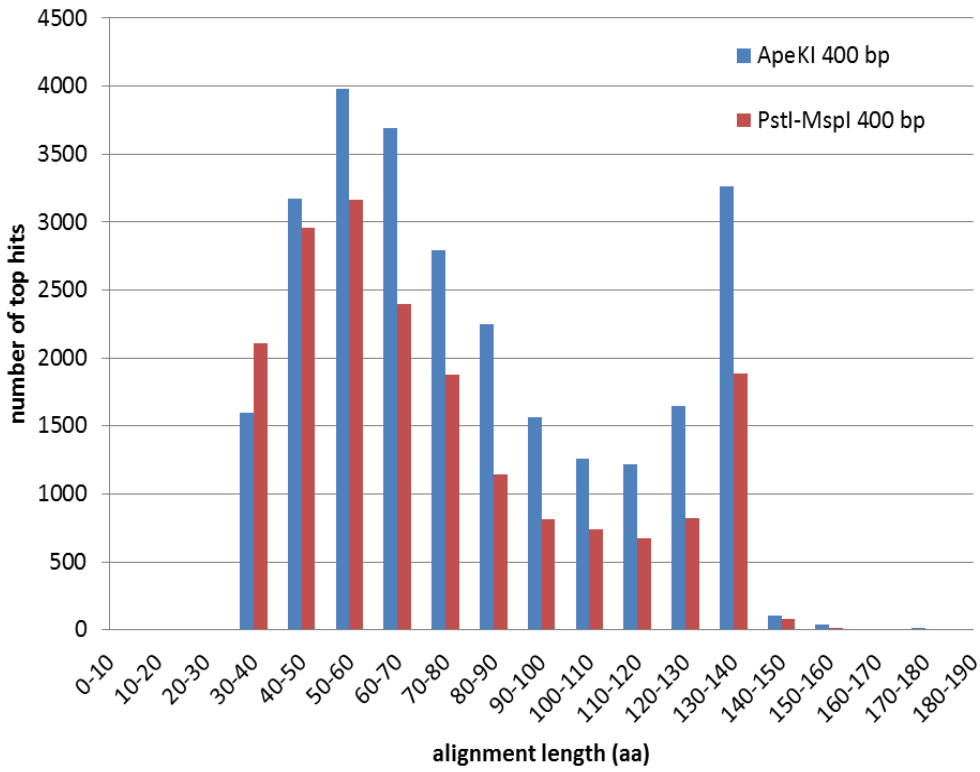


Figure 3-4. Distribution of alignment lengths of BLASTx top hits to ApeKI and PstI-MspI SNP 400 bp sequences. Both apomorphic and non-apomorphic SNP sequences are included. Top hits beyond the maximum alignment length of 133.3 aa are from gaps in the alignment.

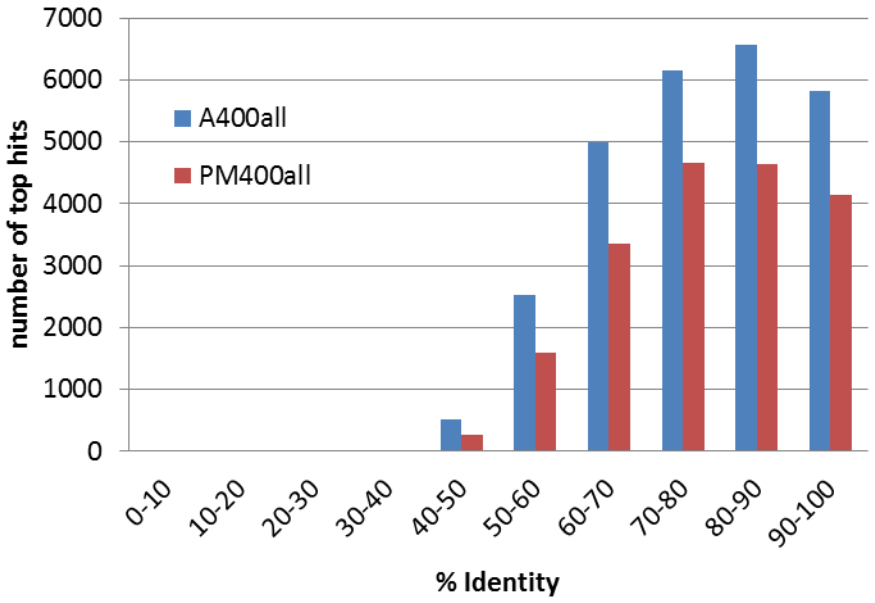


Figure 3-5. Distribution of percent identity of the alignment length (not the query length) of ApeKI (mean 78.10%) and PstI-MspI (mean 78.62%) BLASTx top hits for all 400 bp SNP sequences. Both apomorphic and non-apomorphic SNP sequences are included.

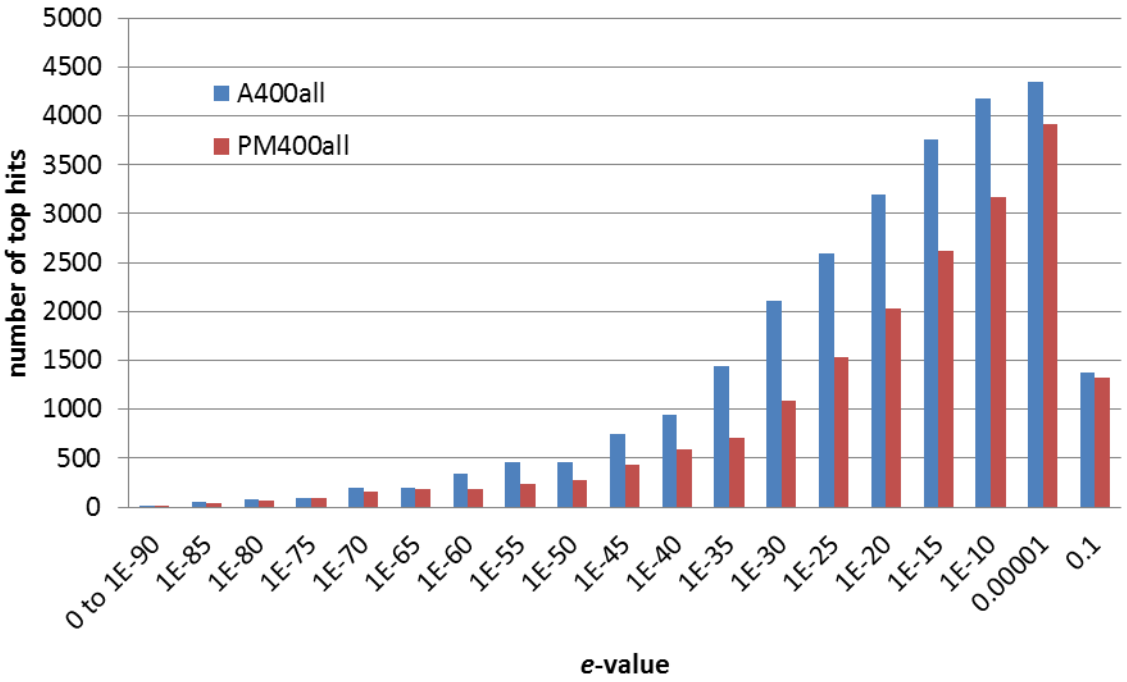


Figure 3-6. Distribution of *e*-values of all ApeKI (mean 9.08E-06) and PstI-MspI (mean 1.25E-05) BLASTx top hits for all 400 bp SNP sequences. Both apomorphic and non-apomorphic SNP sequences are included.

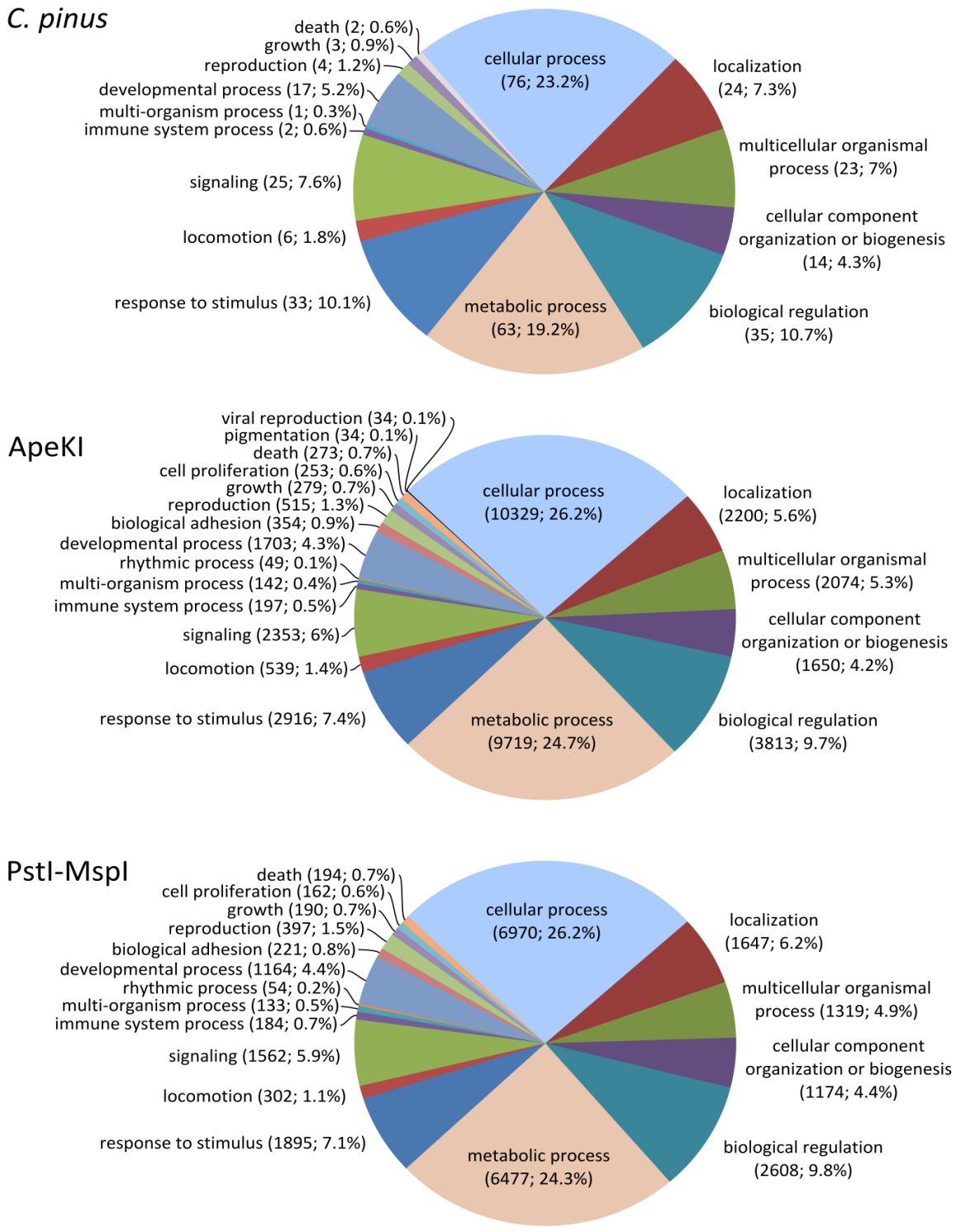


Figure 3-7. Proportions of sequences associated with biological processes at gene ontology level 2 for *C. pinus*, all ApeKI SNPs, and all PstI-MspI SNPs. The number of sequences and their percent are in brackets. A 20 sequence filter cut-off was used for ApeKI and PstI-MspI.

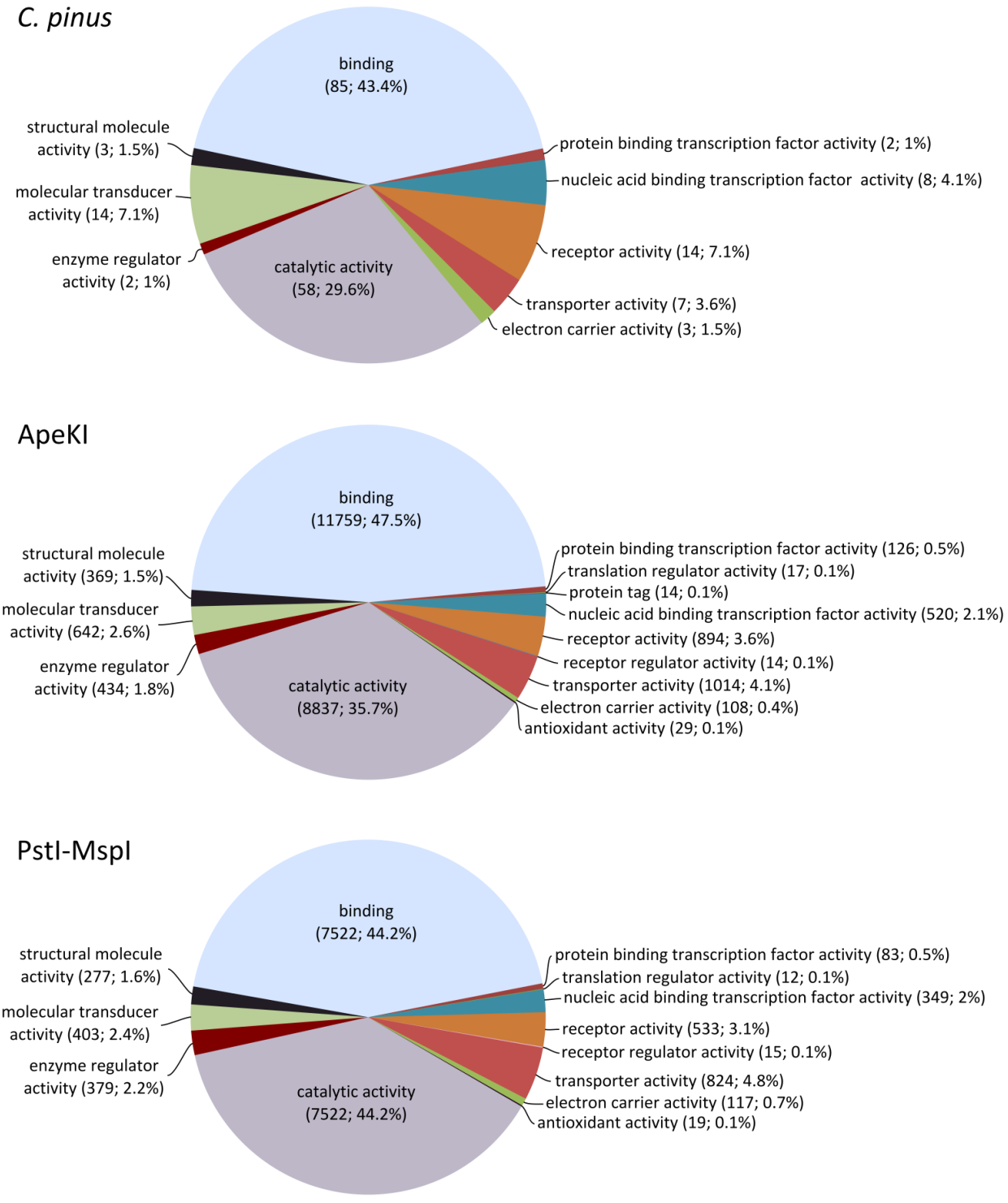


Figure 3-8. Proportions of sequences associated with molecular functions at gene ontology level 2 for *C. pinus*, all ApeKI SNPs, and all PstI-MspI SNPs. The number of sequences and their percent are in brackets. A 5 sequence filter cut-off was used for ApeKI and PstI-MspI.

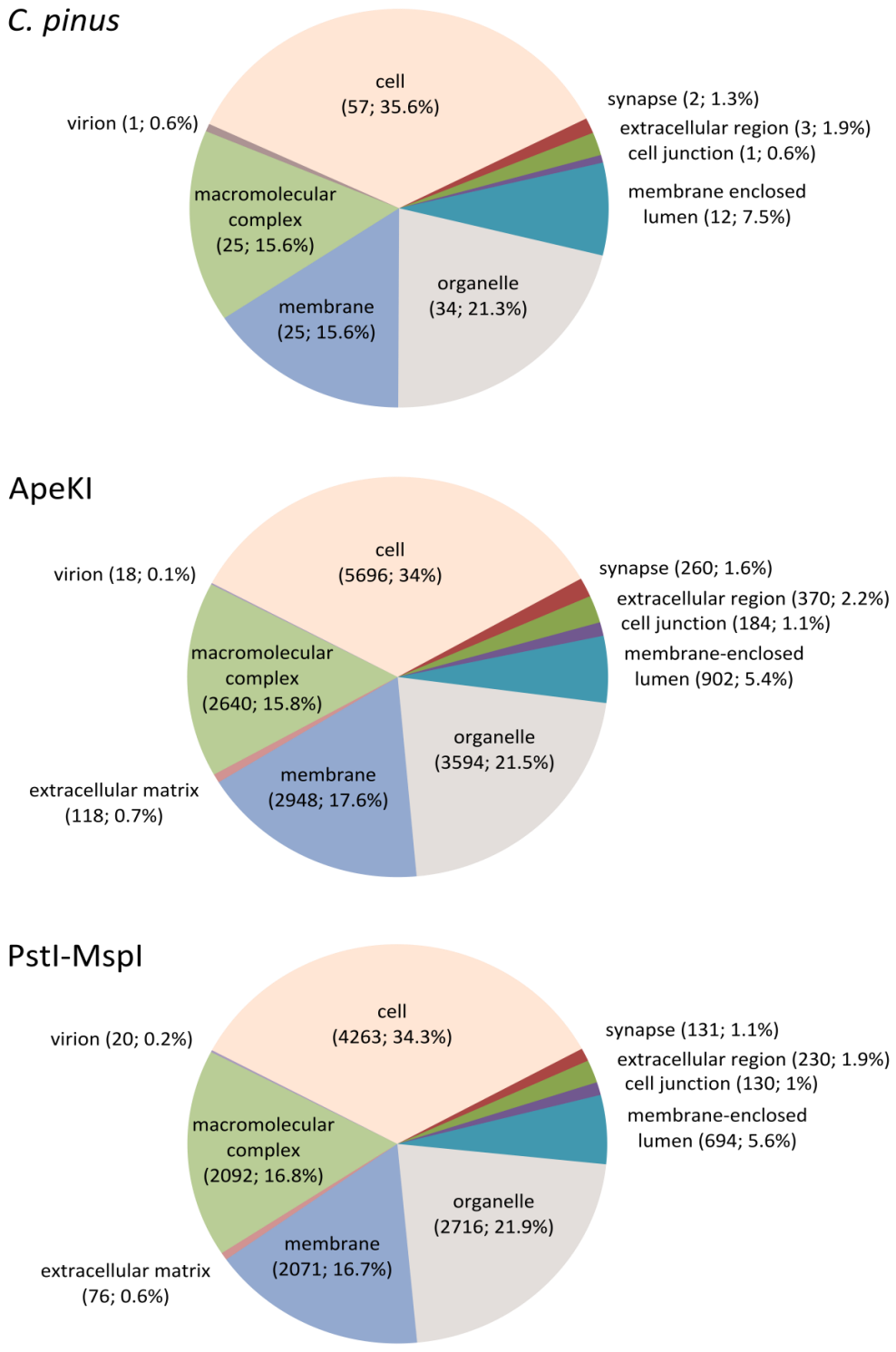


Figure 3-9. Proportions of sequences associated with cellular components at gene ontology level 2 for *C. pinus*, all ApeKI SNPs, and all PstI-MspI SNPs. The number of sequences and their percent are in brackets. A 5 sequence filter cut-off was used for ApeKI and PstI-MspI.

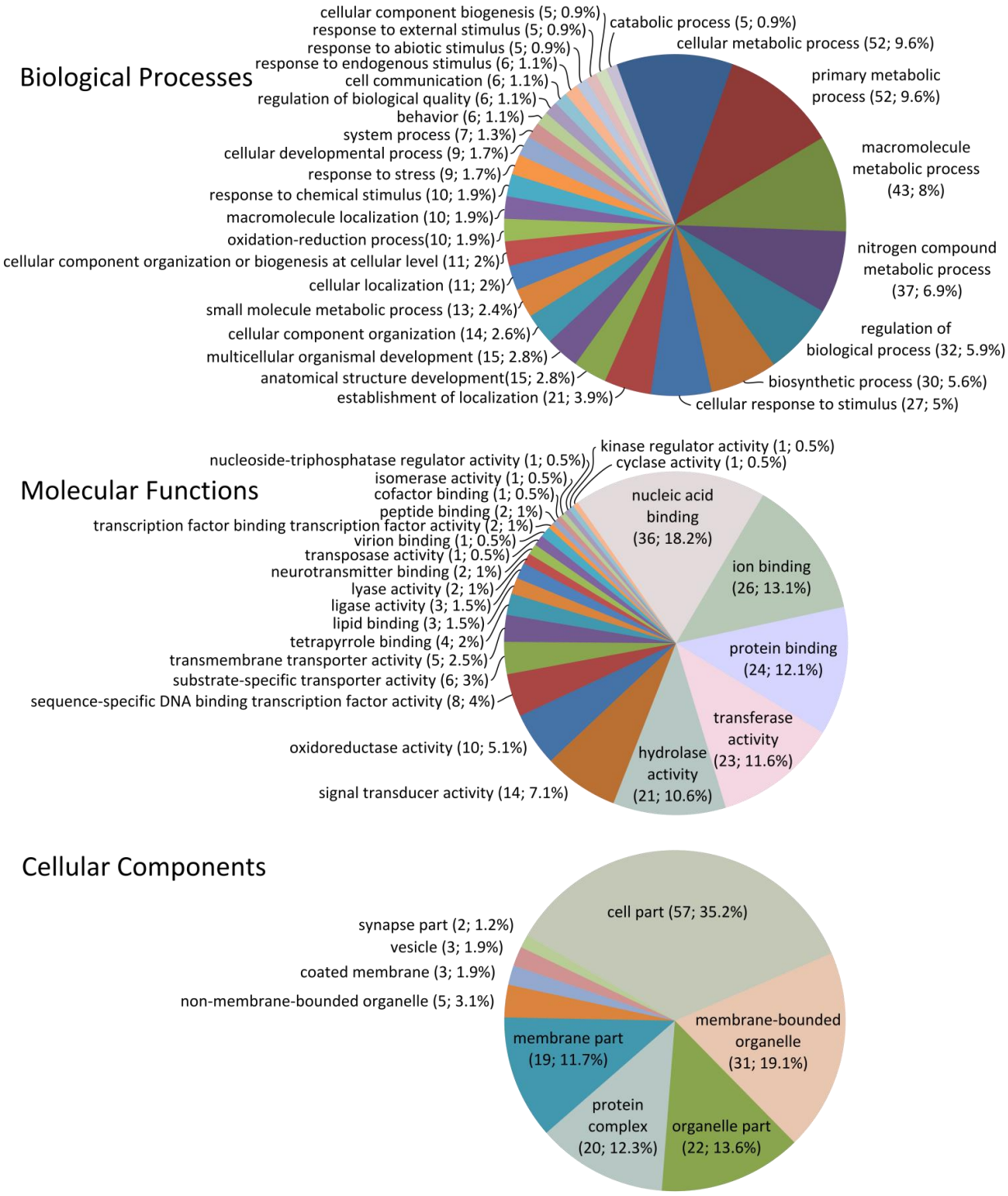


Figure 3-10. Proportions of sequences associated with biological processes, molecular functions, and cellular components at gene ontology level 3 containing autapomorphic SNPs for *C. pinus*. The number of sequences and their percent are in brackets. A 5 sequence filter cut-off was used for biological processes.

3.5. References

Abe, H., Mita, K., Yasukochi, Y., Oshiki, T., Shimada, T. 2005. Retrotransposable elements on the W chromosome of the silkworm, *Bombyx mori*. Cytogenet. Genome Res. 110, 144-151.

Adams, M.D., Celniker, S.E., Holt, R.A. et al., (196 co-authors). 2000. The Genome sequence of *Drosophila melanogaster.* Science. 287. 2185-2195.

Andersson, L. 1990. The driving force: species concepts and ecology. Taxon. 39, 375–382.

Arnold, B., Corbett-Detif, R.B., Hartl, D., Bomblies, K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. 22, 3179-3190.

Bengtsson, J.M., Trona, F., Montagné, N., Anfora, G., Ignell, R., Witzgall, P., Jacquin-Jolly, E. 2012. Putative chemosensory receptors of the codling moth, *Cydia pomonella*, identified by antennal transcriptome analysis. PLoS ONE. 7, e31620.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I, Lipman, D.J., Ostell, J., Sayers, E.W. 2013. GenBank. Nucleic Acids Res. (Database issue): D36-42.

Bissinger, B.W., Donohue, K.V., Khalil, S.M.S., Grozinger, C.M., Sonenshine, D.E., Zhu, J., Roe, R.M. 2011. Synganglion transcriptome and developmental global gene expression in adult females of the American dog tick, *Dermacentor variabilis* (Acari: Ixodidae). Insect Mol. Biol. 20, 465-491.

Bodily, K.D., Morrison, C.M., Renden, R.B., Broadie, K. 2001. A novel member of the Ig superfamily, turtle, is a CNS-specific protein required for coordinated motor control. J. Neurosci. 21, 3113-3125.

Bonin, A., Paris, M., Tetreau, G., David, J.-P., Després, L. 2009. Candidate genes revealed by a genome scan for mosquito resistance to a bacterial insecticide: sequence and gene expression variations. BMC Genomics. 10, 551.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 23, 2633-2635.

Briscoe, A.D., Macias-Muñoz, A., Kozak, K.M., Walters, J.R., Yuan, F., Jamie, G.A., Martin, S.H., Dasmahapatra, K.K., Ferguson, L.C., Mallet, J., Jacquin-Jolly, E., Jiggins, C.D. 2013. Female behaviour drives expression and evolution of gustatory receptors in butterflies. PLoS Genet. 9, e1003620.

Brown, J.W., Baixeras, J., Brown, R., Horak, M., Komai, F., Metzler, E., Razowski, J., Tuck, K. 2005. World catalogue of insects - Tortricidae (Lepidoptera). Apollo Books, Stenstrup, Denmark.

Chang, N.-S., Doherty, J., Ensign, A., Schultz, L., Hsu, L.-J., Hong, Q. 2005. WOX1 is essential for tumor necrosis factor-, UV light-, staurosporine-, and p53-mediated cell death, and its tyrosine 33-phosphorylated form binds and stabilizes serine 46-phosphorylated p53. J. Biol. Chem. 208, 43100-43108.

Chapman, T., Davies, S.J. 2004. Function and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides. 25, 1477-1490.

Chow, L.M.L., Fournel, M., Davidson, D., Veillette, A. 1993. Negative regulation of T-cell receptor signalling by tyrosine protein kinase p50^{csk}. Nature. 365, 156-160.

Chown, A.L., Gaston, K.J. 1999. Exploring links between physiology and ecology at macro-scales: the role of respiratory metabolism in insects. Biol. Rev. Camb. Philos. Soc. 74, 87-120.

Christiansen, J., Kolt, A.M., Hansen, T.V., Nielsen, F.C. 2009. IGF2 mRNA-binding protein 2: biological function and putative role in type 2 diabetes. J. Mol. Endocrinol. 43, 187-195.

Chui, D.S., Oram, J.F., LeBoeuf, R.C., Alpers, C.E., O-Brien, K.D. 1997. High-density lipoprotein-binding protein (HBP)/vigilin is expressed in human

atherosclerotic lesions and colocalizes with apolipoprotein E. Arterioscler. Thromb. Vasc. Biol. 17, 2350-2358.

Conesa, A., Götz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21, 3674-3676.

Cracraft, J. 2000. Species concepts in theoretical and applied biology: A systematic debate with consequences. In: Wheeler, Q.D., Meier, R. (Eds.), Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, USA, pp. 6-7.

Dale, J.W., von Schantz, M., Plant, N. 2012. From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition, John Wiley & Sons, Ltd., University of Surrey, Surrey, UK, pp. 275-303.

Das, B.K., Xia, L., Palandjian, L., Gozani, O., Chyung, Y., Reed, R. 1999. Characterization of a protein complex containing spliceosomal proteins SAPs 49, 130, 145, and 155. Mol. Cell. Biol. 19, 6796-6802.

De Queiroz, K. 2007. Species concepts and species delimitation. Syst. Biol. 56, 879-886.

De Queiroz, K. 2011. Branches in the lines of descent: Charles Darwin and the evolution of the species concept. Biol. J. Linnean Soc. 103, 19-35.

Doran, J.F., Jackson, P., Kynoch, P.A.M., Thompson, P.J. 1983. Isolation of PGP 9.5, a new human neurone-specific protein detected by high-resolution two-dimensional electrophoresis. J. Neurochem. 40, 1542-1547.

Fischer, H.M., Wheat, C.W., Heckel, D.G., Vogel, H. 2008. Evolutionary origins of a novel host plant detoxification gene in butterflies. Mol. Biol. Evol. 25, 809–820.

Fitzpatrick, M.L., Ben-Shahar, Y., Smid, H.M., Vet, L.E.M., Robinson, G.E., Sokolowski, M.B. 2005. Candidate genes for behavioural ecology. Trends Ecol. Evol. 20, 96-104.

Fleischer, T.C., Yun, U.J., Ayer, D.E. 2003. Identification and characterization of three new components of the mSin3A corepressor complex. Mol. Cell. Biol. 23, 3456-3467.

Fordyce, J.A., Nice, C.C., Forister, M.L., Shapiro, A.M. 2002. The significance of wing pattern diversity in the Lycaenidae: mate discrimination by two recently diverged species. J. Evol. Biol. 15, 871-879.

Freeman, T.N. 1953. The spruce budworm, *Choristoneura fumiferana* (Clem.) and an allied new species on pine (Lepidoptera: Tortricidae). Can. Entomol. 85, 121-127.

Freeman, T.N. 1967. On coniferophagous species of *Choristoneura* (Lepidoptera: Tortricidae) in North America I. Some new forms of *Choristoneura* allied to *C. fumiferana*. Can. Entomol. 99, 449-455.

Freeman, T.N., Stehr, G.W. 1967. On coniferophagous species of *Choristoneura* (Lepidoptera: Tortricidae) in North America VI. A summary of the preceding five papers. Can. Entomol. 99, 504-506.

Glaubitz, J., Harriman, J., Casstevens, T. 2012. TASSEL 3.0 Genotyping By Sequencing (GBS) pipeline documentation. Available from www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf. Accessed 2012-2-24.

González-Martínez, S.C., Ersoz, E., Brown, G.R., Wheeler, N.C., Neale, D.B. 2006. DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics. 172, 1915-1926.

Götz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J., Conesa, A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic. Acids. Res. 36, 3420-3435.

Graham, S.M., Watt, W.B., Gall, L.F. 1980. Metabolic resource allocation vs. mating attractiveness: adaptive pressures on the “alba” polymorphism of *Colias* butterflies. Proc. Natl. Acad. Sci. USA. 77, 3615-3619.

Griffith, L.C., Wang, J., Zhong, Y., Wo, C.-F., Greenspan, R.J. 1994. Calcium/calmodulin-dependent protein kinase II and potassium channel subunit Eag similarly affect plasticity in *Drosophila*. Neurobiology. 91, 10044-10048.

Haldar, S.M., Jeyaraj, D., Anand, P., Zhu, H., Lu, Y., Prosdocimo, D.A., Eapen, B., Kawanami, D., Okutsu, M., Brotto, L., Fujioka, H., Kerner, J., Rosca, M.G., McGuinness, O.P., Snow, R.J., Russell, A.P., Gerber, A.N., Bai, X., Yan, Z., Nosek, T.M., Brotto, M., Hoppel, C.L., Jain, M.K. 2012. Kruppel-like factor 15 regulates skeletal muscle lipid flux and exercise adaptation, Proc. Natl. Acad. Sci. USA. 109, 6739–6744.

Han, E.-N., Bauce, E. 1998. Timing of diapause initiation, metabolic changes and overwintering survival of the spruce budworm, *Choristoneura fumiferana*. Ecol. Entomol. 23, 160-167.

Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G., Rienzo, A.D. 2008. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet. 4, e32.

Harvey, G.T. 1996. Genetic relationships among *Choristoneura* species (Lepidoptera: Tortricidae) in North America as revealed by isozyme studies. Can. Entomol. 128, 245-262.

Hennig, W. 1966. Phylogenetic Systematics. University of Illinois Press, Urbana, IL, USA.

Higginbottom, A., Takahashi, Y., Bolling, L., Coonrod, S.A., White, J.M., Partridge, L.J., Monk, P.N. 2003. Structural requirements for the inhibitory action of the CD9 large extracellular domain in sperm/oocyte binding and fusion. Biochem. Biophys. Res. Commun. 311, 208-214.

Homyk, T., Sheppard, D.E. 1977. Behavioral mutants of *Drosophila melanogaster*. I. Isolation and mapping of mutations which decrease flight ability. Genetics. 87, 95-104.

Hufford, M.B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chai, J.-M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C., Guill, K.E., Kaeppler, S.M., Lai, J., Morrell, P.L., Shannoon, L.M., Song, C., Springer, N.M., Swanson-Wagner, R.A., Tiffin, P. Wang, J., Zhang, G., Doebley, J., McMullen, M.D., Ware, D., Buckler, E.S., Yang, S., Ross-Ibarra, J. 2012. Comparative population genomics of maize domestication and improvement. Nature Genet. 44, 808-811.

Hughes, I., Blasiole, B., Huss, D., Warchol, M. E., Rath, N. P., Hurle, B., Ignatova, E., Dickman, J.D., Thalmann, R., Levenson, R., Ornitz, D.M.. 2004. Otopetrin 1 is required for otolith formation in the zebrafish *Danio rerio*. Dev. Biol. 276, 391–402.

Igakura, T., Kadomatsu, K., Taguchi, O., Muramatsu, H., Kaname, T., Miyauchi, T., Yamamura, K., Arimura, K., Muramatsu, T. 1996. Roles of basigin, a member of the immunoglobulin superfamily, in behavior as to an irritating odor, lymphocyte response, and blood–brain barrier. Biochem. Biophys. Res. Commun. 224, 33-36.

Ishihara, M., Shimada, M. 1995. Trade-off in allocation of metabolic reserves: effects of diapause on egg production and adult longevity in a multivotine bruchid, *Kytorhinus sharpianus*. Funct. Ecol. 9, 618-624.

Izadi, H., Samih, M.A., Behroozzy, E., Hadavi, F., Mahdian, K. 2011. Energy allocation changes during diapause in overwintering larvae of pistachio twig borer, *Kermania pistaciella* Amsel (Lepidoptera: Tineidae) in Rafsanjan. ARPN J. Agr. Biol. Sci. 6, 12-17.

Jones, C.D. 1998. The genetic basis of *Drosophila sechellia*'s resistance to a host plant toxin. Genetics. 149, 1899-1908.

Kawaoka, S., Kadota, K., Arai, Y., Suzuki, Y., Fujii, T., Abe, H., Yasukochi, Y., Mita, K., Sugano, S., Shimizu, K., Tomari, Y., Shimada, T., Katsuma, S. 2011. The silkworm W chromosome, is a source of female-enriched piRNAs. RNA. 17, 2144-2151.

Larson, E.L., Andrés, J.A., Bogdanowicz, S.M., Harrison, R.G. 2013. Differential introgression in a mosaic hybrid zone reveals candidate barrier genes. Evolution. Pre-pub online version http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1558-5646/accepted

Lassance, J.-M., Groot, A.T., Liénard, M.A., Antony, B., Borgwardt, C., Andersson, F., Hedenström, E., Heckel, D.G., Löfstedt, C. 2010. Allelic variation in a fatty-acyl reductase gene causes divergence in moth sex pheromones. Nature. 466, 486-489.

Leary, G.P., Allen, J.E., Bunger, P.L., Luginbill, J.B., Linn, C.E., Macallister, I.E., Kavanaugh, M.P., Wanner, K.W. 2012. Single mutation to a sex pheromone receptor provides adaptive specificity between closely related moth species. Proc. Natl. Acad. Sci. USA. 109, 14081-14086.

Lee, C.E. 2002. Evolutionary genetics of invasive species. Trends Ecol. Evol. 17, 386-391.

Lee, E.-C., Yu, S.-Y., Baker, N.E. 2000. The Scabrous protein can act as an extracellular antagonist of Notch signaling in the *Drosophila* wing. Curr. Biol. 10, 931-934.

Li, H., Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 25, 1754-60.

Lindebro, M.C., Poellinger, L., Whitelaw, M.L. 1995. Protein-protein interaction via PAS domains: role of the PAS domain in positive and negative regulation of the bHLH/PAS dioxin receptor-Arnt transcription factor complex. EMBO J. 14, 3528-3539.

Lindroth, R.L. 1991. Biochemical ecology of Aspen-Lepidoptera interactions. J. Kans. Entomol. Soc. 64, 372-380.

Lumley, L.M. Sperling, F.A.H. 2010. Integrating morphology and mitochondrial DNA for species delimitation within the spruce budworm (*Choristoneura fumiferana*) cryptic species complex (Lepidoptera: Tortricidae). Syst. Entomol. 35, 416-428.

Lumley, L.M., Sperling, F.A.H. 2011. Utility of microsatellites and mitochondrial DNA for species delimitation in the spruce budworm (*Choristoneura fumiferana*) species complex (Lepidoptera: Tortricidae). Mol. Phylogenet. Evol. 58, 232-243.

Martinek, S., Inonog, S., Manoukian, A.S., Young, M.W. 2001. A role for the segment polarity gene shaggy/GSK-3 I the *Drosophila* circadian clock. Cell. 105, 769-779.

Mayr, E. 1942. Systematics and the origin of species. Columbia University Press, New York, NY, USA.

Mayr, E. 1982. The Growth of Biological Thought: Diversity, Evolution, and Inheritance. Belknap Press, Cambridge, MA, USA.

Mayr, E. 2000. The Biological Species Concept. In: Wheeler, Q.D., Meier, R. (Eds.), Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, USA, pp. 17-29.

McMahan, U.J., Horton, S.E., Werle, M.J., Honig, L.S., Kröger, S., Ruegg, M.A., Esher, G. 1992. Agrin isoforms and their role in synaptogenesis. Curr. Opin. Cell Biol. 4, 869-874.

Meier, R. 2008. DNA sequences in taxonomy, opportunities and challenges. In: Wheeler, Q.D. (Ed.), The New Taxonomy. CRC Press Taylor & Francis Group, Boca Raton, FL, USA, pp. 95-127.

Meier, R., Willmann, R. 2000. The Hennigian Species Concept. In: Wheeler, Q.D., Meier, R. (Eds.), Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, USA, pp. 30-43.

Meyerson, M., Enders, G.H., Wu, C.L., Su, L.K., Gorka, C., Nelson, C. Harlow, E., Tsai, L.H. 1992. A family of human cdc2-related protein kinases. EMBO. J. 11, 2909-2917.

Mimura, J., Ema, M., Sogawa, K., Fujii-Kuriyama, Y. 1999. Identification of a novel mechanism of regulation of Ah (dioxin) receptor function. Genes Dev. 13, 20-25.

Mishler, B.D., Theriot, E.C. 2000. The Phylogenetic Species Concept (*sensu* Mishler and Theriot). In: Wheeler, Q.D., Meier, R. (Eds.), Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, USA, pp. 44-54.

NCBI. 2013. Help pages, frequently asked qestions. Available from http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ. Access 2013-09-17.

Newton, A.C., Watt, A.D., Lopez, F., Cornelius, J.P., Mesén, J.F., Corea, E.A. 1999. Genetic variation in host susceptibility to attack by the mahogany shoot borer, *Hypsipyla grandella* (Zeller). Agric. For. Entomol. 1, 11-18.

Noack, C., Zafiriou, M.P., Schaeffer, H.J., Rengerm A., Pavlova, E., Dietz, R., Zimmermann, W.H., Bergmann, M.W., Zelarayán, L.C. 2012. Krueppel-like factor 15 regulates Wnt/β-catenin transcription and controls cardiac progenitor cell fate in the postnatal heart. EMBO Mol. Med. 4, 992-1007.

Nosil, P. 2012. Ecological Speciation. Oxford series in ecology and evolution. In: Harvey, P.H., May, R.M., Godfrey, C.J., Dunne, J.A. (Eds.), Oxford University Press, Oxford, U.K.

Olson, M.V., Varki, A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. Nature Reviews Genetics. 4, 20-28.

Powell, J.A. 1995. Biosystematic Studies of Conifer-feeding *Choristoneura* (Lepidoptera: Tortricidae) in the Western United States. University of California Press, Berkeley, CA, USA.

Powell, J.A., De Benedictus, J.A. 1995. Evolutionary interpretation, taxonomy and nomenclature. In: Powell, J.A. (Ed.), Biosystematic Studies of Conifer-feeding *Choristoneura* (Lepidoptera: Tortricidae) in the Western United States. University of California Press, Berkeley, CA, USA, pp.219–275.

Pullin, A.S. 1987. Adult feeding time, lipid accumulation, and overwintering in *Aglais urticae* and *Inachis io* (Lepidoptera: Nymphalidae). J. Zool. 211, 631-641.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R. 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33 (Web Server issue), W116-W120.

Ralph, S., Yueh, H., Friedmann, M., Aeschliman, D., Zeznik, J.A., Nelson, C.C., Butterfield, Y.S.N., Kirkpatrick, R., Liu, J., Jones, S.J.M., Marra, M.A. Douglas, C.J., Ritland, K., Bohlmann, J. 2006. Conifer defence against - insects, microarray gene expression profiling of Sitka spruce (*Picea sitchensis*) induced by mechanical wounding or feeding by spruce budworms (*Choristoneura occidentalis*) or white pine weevils (*Pissodes strobi*) reveals large-scale changes of the host transcriptome. Plant Cell Environ. 29, 1545-1570.

Sanders, C.J., Daterman, G.E., Ennis, T.J. 1977. Sex pheromone responses of *Choristoneura* spp. and their hybrids (Lepidoptera: Tortricidae). Can. Ent. 109, 1203-1220.

Sano, Y., Akimaru, H., Okamura, T., Nagao, T., Okada, M., Ishii, S. 2005. *Drosophila* activating transcription factor-2 is involved in stress response via activation by p38, but not c-Jun NH_2-terminal kinase. Mol. Biol. Cell. 16, 2934-2946.

Schaack, S., Gilbert, C., Feschotte, C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in Ecol. Evol. 25, 537-546.

Schuler, M.A. 1996. The role of cytochrome P450 monooxygenases in plant-insect interactions. Plant Physiol. 112, 1411-1419.

Scott, J.G., Liu, N., Wen, Z. 1998. Insect cytochromes P450: diversity, insecticide resistance and tolerance to plant toxins. Comp. Biochem. Physiol. C. 121, 147-155.

Scott, J.G., Wen, Z. 2001. Cytochromes P450 of insects: the tip of the iceberg. Pest Manag. Sci. 57, 958-967.

Silk, P.J., Kuenen, L.P.S. 1988. Sex pheromones and behavioral biology of the coniferophagous *Choristoneura*. Ann. Rev. Entomol. 33, 83-101.

Silva, J.C., Potts, B.M., Bijma, P., Kerr, R.J., Pilbeam, D.J. 2013. Genetic control of interactions among individuals: contrasting outcomes of indirect genetic effects arising from neighbour disease infection and competition in a forest tree. New Phytologist. 197, 631-641.

Simpson, G.G. 1951. The species concept. Evolution. 5, 285–298.

Sperling, F.A.H. 2003. Butterfly molecular systematics: from species definitions to higher-level phylogenies. In: Boggs, C.L., Watt, W.B., Ehrilch, P.R. (Eds.), Butterflies: Ecology and Evolution Taking Flight, University of Chicago Press, Chicago, IL, USA, pp. 431-458.

Sperling, F.A.H., Hickey, D.A. 1994. Mitochondrial DNA sequence variation in the spruce budworm species complex (*Choristoneura*: Lepidoptera) Mol. Biol. Evol. 11, 656-665.

Sperling, F.A.H., Hickey, D.A. 1995. Amplified mitochondrial DNA as a diagnostic marker for species of conifer-feeding *Choristoneura* (Lepidoptera: Tortricidae). Can. Entomol. 127, 277-288.

Templeton, A.R. 1989. The meaning of species and speciation: a genetic perspective. In: Otte, D., Endler, J.A. (Eds.), Speciation and its consequences, Sinauer, Sunderland, MA, USA, pp. 159-183.

Terriente-Félix, A., de Celis, J.F. 2009. Osa, a subunit of the BAP chromatin-remodelling complex, participates in the regulation of gene expression in response to EGFR signalling in the *Drosophila* wing. Dev. Biol. 392, 350-361.

True, J.R. 2003. Insect melanism: the molecules matter. Trends Ecol. Evol. 18, 640-647.

Vakenti, J.M., Cossentine, J.E., Cooper, B.E., Sharkey, M.J., Yoshimoto, C.M., Jensen, L.B.M. 2001. Host-plant range and parasitoids of obliquebanded and three-lined leafrollers (Lepidoptera: Tortricidae). Can. Entomol. 133, 139-146.

Van Valen, L. 1976. Ecological species, multispecies, and oaks. Taxon 25, 233–239.

Vogel, H., Altincicek, B., Glöckner, G., Vilcinskas, A. 2011. A comprehensive transcriptome and immune-gene repertoire of the lepidopteran model host *Galleria mellonella*. BMC Genomics. 12, 308.

Volney, W.J.A., Fleming, R.A. 2007. Spruce budworm (*Choristoneura* spp.) biotype reactions to forest and climate characteristics. Glob. Change Biol. 13, 1630-1643.

Wallin, K.F., Raffa, K.F. 1999. Altered constitutive and inducible phloem monoterpenes following natural defoliation of jack pine: Implications to host mediated interguild interactions and plant defense theories. J. Chem. Ecol. 25, 861-880

Wanner, K.W., Nichols, A.S., Allen, J.E., Bunger, P.L., Garczynski, S.F., Linn, C.E., Robertson, H.M., Luetje, C.W. 2010. Sex pheromone receptor

specificity in the European corn borer moth, *Ostrinia nubilalis*. PLoS ONE. 5, e8685.

Warnecke, M., Oster, H., Revelli, J.P. Alvarez-Bolado, G., Eichele, G. 2005. Abnormal development of the locus coeruleus in Ear2(Nr2f6)-deficient mice impairs the functionality of the forebrain clock and affects nociception. Genes Dev. 19, 614-625.

Wheeler, Q.D., Platnick, N.I. 2000. The Phylogenetic Species Concept (*sensu* Wheeler and Platnick). In: Wheeler, Q.D., Meier, R. (Eds.), Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, USA, pp. 55-69.

Whitelaw, M., Pongratz, I., Wilhelmsson, A., Gustafsson, J.A., Poellinger, L. 1993. Ligand-dependent recruitment of the Arnt coregulator determins DNA recognition by the dioxin receptor. Mol. Cell. Biol. 13, 2504-2514.

Wiley, E.O., Mayden, R.L. 2000. The Evolutionary Species Concept. In: Wheeler, Q.D., Meier, R. (Eds.), Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, USA, pp. 70-89.

Wu, S.F., Yao, Y., Huang, J., Ye, G.Y. 2012. Characterization of a β-adrenergic-like octopamine receptor from the rice stem borer (*Chilo suppressalis*). J. Exp. Biol. 215, 2646-2652.

Xu, Y., Malhotra, A., Ren, M., Schlame, M. 2006. The enzymatic function of tafazzin. J. Biol. Chem. 281, 39217-39224.

Yoon, C. K. 2010. Naming Nature. W.W. Norton & Company, Inc., New York, NY, USA.

Yuan, Z., Wagner, L., Poloumienko, A., Bakovic, M. 2004. Identification and expression of a mouse muscle-specific CTL1 gene. Gene. 341, 305-312.

Yun, B., Farkas, R., Lee., K., Rabinow, L. 1994. The Doa locus encodes a member of a new protein kinase family and is essential for eye and embryonic development in *Drosophila melanogaster*. Genes Dev. 8, 1160-1173.

Zhu, Y. C., Guo, Z., Chen, M.-S., Zhu, K.Y., Liu, X.F., Scheffler, B. 2011. Major putative pesticide receptors, detoxification enzymes, and transcriptional

profile of the midgut of the tobacco budworm, *Heliothis virescens* (Lepidoptera: Noctuidae). Journal of Invertebr. Pathol. 106, 296-307.

Chapter 4

General Conclusions

4.1. Thesis summary

The objectives of this thesis were to produce a well resolved and accurate species phylogeny of the SBW complex using genome-wide markers, determine the extent of divergences within the complex, compare the usefulness of different restriction enzymes used in genotyping-by-sequencing, find diagnostic or apomorphic SNPs for each species, and gain insight into the kinds of genes that were involved in species divergence, with a focus on *C. pinus.*

Spruce budworm species are closely related and difficult to distinguish using traditional taxonomy (Nealis, 2008). To determine the species phylogeny of this group, applying the multitude of markers produced by next generation sequencing is particularly appropriate. For Chapter 2 we sequenced DNA associated with ApeKI restriction enzyme sites in 102 specimens, and DNA associated with PstI-MspI sites in 144 specimens. This produced over a billion reads total, covering over 16 million unique sequences total with a read depth of two or more. Mining these sequences produced more than one million SNP loci with low genotype coverage, or more than 200 thousand loci that were genotyped in at least 75% of specimens. Using the high coverage SNPs, we found between 4 and 945 unique loci for each of the SBW species, *C. pinus, C. fumiferana, C. occidentalis, C. biennis, C. carnana,* and *C. retiniana*. We also found that, although the restriction enzyme ApeKI cut more frequently and produced roughly four times more SNP loci and greater coverage of the genome, phylogenetic analysis of its associated SNPs produced the same topology as the less frequently cutting PstI-MspI enzyme. All species were monophyletic except *C. occidentalis* and *C. biennis,* and the western species were more closely related to *C. fumiferana* than *C. pinus* was, despite sympatry of the eastern species *C. fumiferana* and *C. pinus*.

Speciation is a fundamental process of evolution. Understanding the genetic processes enabling speciation potentially allows us to understand the mechanisms that produce species-specific morphology, host plant and oviposition substrate preferences, seasonal phenology or daily periodicity of flight time, species distribution, and almost all aspects of their ecology. It also reveals the evolutionary history of the genes enabling these mechanisms of speciation within the SBW lineages. In Chapter 3 we sought to identify genes in the sequence surrounding SNPs with diagnostic genotypes for each species. The species with the largest number of diagnostic SNPs was *C. pinus* and almost a third of the sequences surrounding these SNPs aligned with sequences in the NCBI database. The biological functions of these genes included gustatory receptors, odorant receptors, wing disc development, courtship behaviour, sequence-specific gene expression control, metabolism, and detoxification. Almost all of these functions could be related to species divergences in the SBW group.

Early species concepts were based on phenotypic traits such as morphology and life history (Mayr, 1982), but modern species concepts incorporate phylogenetic information and evolutionary history (Wheeler and Meier, 2000; Simpson, 1951). Evidence of the divergence and integrity of a lineage is indicated by the presence and number of diagnostic phenotypic traits. In Chapter 3 we used genotypic evidence to infer phenotypic divergences for species. Although this is the reverse order to how species have traditionally been distinguished, these results supported the same conclusions that traditional methods produced. The more morphologically and ecologically distinct species were also more genetically distinct.

4.2. Practical relevance

Spruce budworm caterpillars defoliate conifers (Volney and Fleming, 2007), and severe defoliation can cause slower growth, top-kill, and sometimes lead to the death of trees (Nealis and Régnière, 2004; McCullough, 2000; Nealis et al., 2003). Both the larger multinational forestry companies and smaller logging

operations rely on growing and harvesting these trees, and serious SBW outbreaks can have financial consequences. Forestry is multi-year investment, and to protect this investment forest management units are monitored for stand structure, weeds, pests, and disease (AESRD, 2012). Monitoring and protecting forests is important economically, because a single severe pest outbreak can impact stands destined for multiple years' harvests.

Forests are managed to minimize impact of insects and disease (AESRD, 2009). When considering the spread of an infestation, the forester needs to know which tree species are infested and if the pest can attack other tree species as well. The forester uses this information to implement appropriate management tactics, whether to apply chemicals, harvest a buffer zone, re-forest the area with a non-host species, or refrain from applying controls (AESRD, 2009). These are practical concerns and to intelligently address them we need to know what the pest species are, which trees they attack, and if, or how quickly, they can adapt to other tree species. Understanding their adaptive variation and the underlying genetic variation of SBW pest species was one goal of this research.

For irruptive pests, outbreaks typically increase the population size of species and extend their current range. When populations invade new environments, new selection pressures may trigger an adaptive response. For example, in recent years the mountain pine beetle has explosively expanded its range beyond previous northern and eastern boundaries (De la Giroday et al., 2012), and threatens to jump from its historical host, lodgepole pine (*Pinus contorta*), to a new host, jack pine (*P. banksiana*) (Cullingham et al., 2011). For SBW, an expanded range could increase geographic overlap with sister species, and increase the potential for hybridization and introgression. Hybrid populations may contain new genetic combinations that can undergo rapid diversification when under divergent selection (Seehausen, 2004). Adaptive diversification can increase the potential to attack novel hosts or attack at different times of the year. Although examples of such hybrid swarms have been found in whitefish (Hudson et al., 2011) and swallowtail butterflies (Mercader et al., 2009), it is unclear if the SBW species have undergone this process in the past or have potential to do so in

the future. We need to continue learning more about these irruptive pests to understand the genetic mechanisms underlying outbreak behaviour.

4.3. Theoretical relevance

Speciation is a continuous process (Nosil, 2012), and, in much the same way, variation within species is often continuous. Variation in a population facilitates adaptation, by allowing some individuals to survive changes in their environment such as a cold spring, a rainy summer, or a local extinction of a preferred host plant. The offspring of the surviving individuals inherit the alleles that produced greater fitness under the new conditions. The variation in a population not only facilitates adaptation, but, in some cases, speciation.

Speciation can occur with a decrease in genetic variation, an increase in genetic variation, or both. A decrease in genetic variation, such as fixation of alleles, can be caused by divergent selection, bottlenecks, or founder events. For example, two closely related New World oriole species diverged after a founder event where the population that became the new species lost the migration distance that the original species had to its breeding grounds (Kondo et al., 2008). Similarly, in Hawaiian crickets the fixation of many quantitative trait loci are associated with divergence in both male song and female acoustic preference and presumably resulted in speciation (Shaw and Lesnick, 2009).

Increases in genetic variation may also accompany speciation due to adaptive radiation into new ecological niches (Schluter, 2000). Genetic variation can increase through hybridization, introgression, gene duplication events, and mutation. For example, orchids have radiated into multiple species and colonized low altitude habitats by evolving a water-conserving photosynthetic pathway from the ancestral state, C_3 photosynthesis (Silvera et al., 2009). Another example of mutations and gene duplications enabling organisms to utilize new ecological niches is the large gene superfamily cytochrome P450s. Cytochrome P450s are enzymes that metabolize xenobiotics, enabling herbivores to denature plant toxins (Berenbaum et al., 1996). Duplications and subsequent mutations of P450s in

swallowtail butterflies have allowed populations to utilize different host plants (Li et al., 2002). Similarly, duplications and mutations in gustatory receptors have allowed female butterflies to be attracted to specific oviposition plants that differ from plants used by other butterfly lineages (Briscoe et al., 2013).

Both gustatory receptor and cytochrome P450 genes have fixed SNPs in at least one of the SBW species, *C. pinus*. It is likely that the speciation events and subsequent divergences that produced the SBW complex involved a combination of gains and losses of genetic variation. Discovering such genes is scientifically relevant because evolutionary biologists in many fields pursue "the genes that matter", which have alleles producing phenotypic variation and divergence among populations (Edwards, 2013; Rockman, 2012). These genes are a flashlight in the dark for evolutionary biologists. Understanding the genetics of speciation not only allows us to better understand speciation as a process but to better define species, subspecies, and biotypes.

4.4. Future research

The results of this thesis can facilitate many more research projects. The genes associated with apomorphic SNPs are good candidates for knock-out gene studies, and for SNP-chip designs that can be used to measure and monitor genetic variation among species and populations. The raw sequence data can be mined for biogeographical, host association, introgression, and hybridization studies. Sub-sets of the sequence data can be used for single species studies, and although isolation by distance was not detected with this data set, if specimens were more evenly spread throughout the species range, this data could still be a promising starting point for measuring fine-scale population structure.

Two SBW species, *C. lambertiana* and *C. orae*, were not included in this study due to the unavailability of suitable DNA. Parallel sequencing of specimens of these species would allow a more comprehensive phylogeny of the SBW complex to be reconstructed. Similarly, including all the subspecies of *C. lambertiana* and the subspecies *C. pinus maritima* would produce a more

complete and fine scale phylogeny of this group. Because *C. lambertiana* is the only other pine-feeding budworm, it raises the question - is it the sister taxon to *C. pinus* or is it more closely related to the other western species? Previous allozyme analyses placed *C. l. subretiniana* in the western lineage and *C. l. ponderosana* with *C. pinus* outside of it (Harvey, 1996), but mtDNA placed both *C. pinus* and *C. lambertiana* in the western lineage (Lumley and Sperling, 2011). If *C. lambertiana* is more closely related to the western lineage, questions about the identity of the ancestral host plant and the number of host switches between spruce-fir feeding and pine feeding could be addressed.

Availability of specimens of *C. lambertiana* would also determine whether the apomorphic SNPs for *C. pinus* distinguish that species alone, the pine-feeding species at large, or have been convergently derived. Further sequencing of all subspecies and of a greater range of *C. biennis* and *C. occidentalis* specimens could allow further evaluation of their status as subspecies, biotypes, or separate species. Sequencing DNA for species in the genus *Choristoneura* beyond the SBW complex would produce a broader phylogeny and resolve some relationships between deciduous feeders and conifer feeders. However, more distantly related species have more variation in restriction enzyme recognition sites. Consequently, genotyping-by-sequencing would produce fewer usable loci and more null alleles when comparing more divergent species (Arnold et al., 2013; Cariou et al., 2013). But testing the limits of this technique would also be interesting, and possibly provide an additional method of measuring divergence.

The SBW complex is an excellent system for studying the genetics of irruptive forest pests, and also for studying the evolution of speciation. Irruptive forest pests go through outbreak cycles and understanding how genetic diversity fluctuates during these cycles could lead to insights in the ability of species to attack new hosts or invade new habitats. It should be possible to track genetic diversity with the massive amounts of sequence data produced by genotyping-by-sequencing technology. Large sequence datasets bring new challenges to data analyses, but allow us to answer questions that we previously did not have the means to answer.

4.5. References

AESRD. 2009. Forest Management Unit F23 and A09 Operating Ground Rules. http://srd.alberta.ca/LandsForests/ForestManagement/documents/ForestManagementUnit-F23AndA09-OperatingGroundRules-Sep2009.pdf. Retrieved August 12, 2013.

AESRD. 2012. Northeast Alberta Operating Ground Rules. http://srd.alberta.ca/LandsForests/ForestManagement/documents/ALPAC-NEastAlta-OperatingGroundRules-Oct2012.pdf. Retrieved August 12, 2013.

Arnold, B., Corbett-Detif, R.B., Hartl, D., Bomblies, K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. 22, 3179-3190.

Berenbaum, M.R., Favret, C., Schuler, M.A. 1996. On Defining "Key Innovations" in an Adaptive Radiation: Cytochrome P450S and Papilionidae. Amer. Nat. 148, S139-S155.

Briscoe, A.D., Macias-Muñoz, A., Kozak, K.M., Walters, J.R., Yuan, F., Jamie, G.A., Martin, S.H., Dasmahapatra, K.K., Ferguson, L.C., Mallet, J., Jacquin-Jolly, E., Jiggins, C.D. 2013. Female behaviour drives expression and evolution of gustatory receptors in butterflies. PLoS Genet. 9, e1003620.

Cariou, M., Duret, L., Charlat, S. 2013. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. Ecol. Evol. 3, 846-852.

Cullingham, C.I., Cooke, J.E.K., Dang, S., Davis, C.S., Cooke, B.J., Coltman, D.W. 2011. Mountain pine beetle host-range expansion threatens the boreal forest. Mol. Ecol. 20, 2157-2171.

De la Giroday, H.-M.C., Carroll, A.L., Aukema, B.H. 2012. Breach of the northern Rocky Mountain geoclimatic barrier: initiation of range expansion by the mountain pine beetle. J. Biogeogr. 39, 1112-1123.

Edwards, S.V. 2013. Next-generation QTL mapping: crowdsourcing SNPs, without pedigrees. Mol. Ecol. 22, 3885-3887.

Harvey, G.T. 1996. Genetic relationships among *Choristoneura* species (Lepidoptera: Tortricidae) in North America as revealed by isozyme studies. Can. Entomol. 128, 245-262.

Hudson, A.G., Vonlanthen, P., Seehausen, O. 2011. Rapid parallel adaptive radiations from a single hybridogenic ancestral population. Proc. R. Soc. B. 278, 58-66.

Kondo, B., Peters, J.L., Rosensteel, B.B., Omland, K.E. 2008. Coalescent analysis of multiple loci support a new route to speciation in birds. Evolution. 62, 1182-1191.

Li, W., Peterson, R.A., Shuler, M.A., Berenbaum, M.R. 2002. *CYP6B* cytochrome P450 monooxygenases from *Papilio canadensis* and *Papilio glaucus*: potential contributions of sequence divergence to host plant associations. Insect Mol. Biol. 11, 543-551.

Lumley, L.M., Sperling, F.A.H. 2011. Utility of microsatellites and mitochondrial DNA for species delimitation in the spruce budworm (*Choristoneura fumiferana*) species complex (Lepidoptera: Tortricidae). Mol. Phylogenet. Evol. 58, 232-243.

Mayr, E. 1982. The Growth of Biological Thought: Diversity, Evolution, and Inheritance. Belknap Press, Cambridge, MA, USA.

McCullough, D.G. 2000. A review of factors affecting the population dynamics of jack pine budworm (*Choristoneura pinus pinus* Freeman). Popul. Ecol. 42, 243-256.

Mercader, R.J., Aardema, M.L., Scriber, J.M. 2009. Hybridization leads to host-use divergence in a polyphagous butterfly sibling species pair. Oecologia. 158, 651-662.

Nealis, V.G. 2008. Spruce budworms, *Choristoneura* Lederer (Lepidoptera: Tortricidae). In: Capinera, J. (Ed.) Encyclopedia of Entomology, second ed. Springer, Dordrecht, Netherlands, pp. 3524-3531.

Nealis, V.G., Magnussen, S., Hopki, A.A. 2003. A lagged, density-dependent relationship between jack pine budworm *Choristoneura pinus pinus* and its host tree, *Pinus banksiana*. Ecol. Entomol. 28, 183-192.

Nealis, V.G., Régnière, J. 2004. Insect-host relationships influencing disturbance by the spruce budworm in a boreal mixedwood forest. Can. J. Forest Res. 34, 1870-1882.

Nosil, P. 2012. Ecological Speciation. Oxford series in ecology and evolution. Harvey, P.H., May, R.M., Godfrey, C.J., Dunne, J.A. (Eds.) Oxford University Press, Oxford, UK.

Rockman, M.V. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution. 66, 1-17.

Schluter, D. 2000. The ecology of adaptive radiation. Oxford University Press, Oxford, UK.

Seehausen, O. 2004. Hybridization and adaptive radiation. Trends Ecol. Evol. 19, 198-207.

Shaw, K.L., Lesnick, S.C. 2009. Genomic linkage of male song and female acoustic preference QTL underlying a rapid species radiation. Proc. Natl. Acad. Sci. USA. 106, 9737-9742.

Silvera, K. Santiago, L.S., Cushman, J.C., Winter, K. 2009. Crassulacean acid metabolism and epiphytism linked to adaptive radiations in the Orchidaceae. Plant Physiol. 149, 1838-1847.

Simpson, G.G. 1951. The species concept. Evolution 5, 285–298.

Wheeler, Q.D., Meier, R. 2000. Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, USA.

Volney, W.J.A., Fleming, R.A. 2007. Spruce budworm (*Choristoneura* spp.) biotype reactions to forest and climate characteristics. Glob. Change Biol. 13, 1630-1643.

Biography

On October 13, 1986, in the middle of a busy harvest season and after a hearty Thanksgiving dinner, Dixie and Art Bird had their second child at the Foothills hospital in Calgary, Alberta. My brother and I grew up on the family farm near Arrowwood, Alberta, and spent much of our childhood immersed in books and exploring the coulee below our house. Our mother taught us the names of the wild flowers and berries, our father taught us how to halter break steers and drive tractors, and both taught us the love of biology which is prevalent among our ancestors.

In 2004, after graduating from high school, I started my B.Sc. at the University of Lethbridge. My first job as an undergraduate was during the rainy summer of 2005, when I treated sloughs for mosquitoes and dodged discarded bombs on the CFB at Suffield. The second summer was spent running PCRs in a genetics lab of an agricultural researcher at the Lethbridge Research Station. The third summer was spent canoeing the rivers of southern Alberta to count male and female cottonwoods for an ecologist. The fourth and final summer was spent mist netting chickadees and woodpeckers in Revelstoke, Alaska, and Waterton for population genetics studies. After finishing my B.Sc., I worked for a fish toxicologist until realizing that population genetics and ecology studies were a lot of fun, and started looking into grad school.

I started my M.Sc. in Felix Sperling's lab at the University of Alberta in September 2010, after a summer of identifying and pinning ichneumonids for his Ph.D. student, Marla. During the last three years, my love of genetics and molecular evolution has grown. I have developed new skills in teaching undergraduate labs, learned much about insects and entomologists, and gained many new experiences. I am grateful for the learning opportunities, support, and challenges provided here, and I am excited for any future opportunities.