The magnitude of the island of genetic differentiation surrounding a strong, prezygotic reproductive isolating barrier

A thesis submitted by

Jessie Martin

in partial fulfillment of the requirements for the degree of

Master of Science

in

Biology

Tufts University

August, 2015

Advisor: Erik Dopman

Abstract:

Numerous studies have observed regions of elevated genetic differentiation (islands) when comparing the genomes of incipient species. The mechanisms that lead to the formation of these islands nor their basic characteristics (length, frequency, distribution) are well understood. One hypothesis is that islands of elevated genetic differentiation form around loci that encode reproductive isolating barriers, but there is currently little empirical evidence to support this thinking. In this thesis, I test this idea using an emerging model of speciation, the pheromone strains of *Ostrinia nubilalis*. I quantify the size of the island that forms around *pgFAR*, the gene responsible for differential pheromone production in the two strains. I observed the island surrounding *pgFAR* to be several hundred kilobases long, several orders of magnitude smaller than some of the islands observed in previous work. This discrepancy in island sizes suggests that there may be multiple different mechanisms responsible for island creation.

Acknowledgements:

First I would like to thank my advisor, Erik Dopman, for welcoming me into his lab. I very much appreciate all of his advice and support along the way. I would also like thank the members of my committee, Mitch McVey and Phil Starks, for their thoughful feedback and recommendations. Thanks as well to everyone in the lab during my time: Genny Kozak, Crista Wadsworth, Nooria Al-Wathiqui , Gabriel Golczer, Rebecca Levy, and our undergraduates: Shoshanna Kahne, Gabriel Spieler, Sebi Zahler, Zoe Greene, and Yuta Okada. I truly appreciate their support and companionship during my time in the lab. Lastly, I would like to thank my husband and parents for all of their love and encouragement throughout the entire process.

Table of Contents:

List of Tables

CHAPTER 2

List of Figures

CHAPTER 2

CHAPTER 3

The magnitude of the island of genetic differentiation surrounding a strong, prezygotic reproductive isolating barrier

CHAPTER 1

Introduction

"Without speciation, there would be no diversification of the organic world, no adaptive radiation, and very little evolutionary progress. The species, then, is the keystone of evolution." – Ernst Mayr, Animal Species and Evolution*, 1963*

In 1859, Charles Darwin published *On the Origin of Species* in which he articulated the theory of evolution by natural selection. Darwin posited that heritable biological traits would become more or less common within a population depending on the degree of reproductive success they conferred on the individuals who possessed them. Overtime, populations would evolve to be better suited to their environments. Ironically, despite the title of the book, Darwin did not focus on describing the process by which new species arise; he primarily addressed the changes that occur along the branches of the phylogenic tree, but not what causes branches to split.

Thus, following the acceptance of Darwin's theory, one of the major tasks facing the field of evolutionary biology was to uncover the processes that lead to the splitting of phylogenic branches. Ernst Mayr tackled this problem by offering a new definition of species: a group of individuals that could breed among themselves, but not with others (Mayr, 1942). Implicit in this definition is the concept of reproductive isolating barriers, traits that prevent individuals from two different species from successfully breeding or which prevent hybrid offspring from breeding with the parental populations.

Mayr contributed much to our understanding of how new species form on the organismal level, but the processes by which divergence happens on the level of the genome have remained something of a mystery. Which genes contribute to speciation? What are their relative effect sizes? How do these speciation genes influence the genome around them? Are important changes more likely to occur in regulatory sites or coding regions? Are the same genes repeatedly involved in different speciation events? (Nosil and Schluter, 2001). These questions are particularly challenging for subspecies that are diverging while still exchanging genes.

With the advent of new sequencing technologies, evolutionary biologists are now able to ask previously unthinkable questions regarding what happens to the genome during the process of speciation. In the past decade, researchers have taken advantage of new technologies to examine the genomes of pairs of species or subspecies and observe the degree to which they are similar and different from each other. These genome scans (e.g. Turner *et al.*, 2005; Ellegren *et al.*, 2012; Phifer-Rixey *et al.*, 2014) have provided evidence in the support of the genic view of speciation (Wu, 2001), which argues that gene flow is heterogeneous across the genome. At loci for reproductive isolating barriers and local adaptation, selection inhibits the gene flow from one population to another, while gene flow can continue uninhibited elsewhere in the genome. This heterogeneous gene flow results in a pattern of alternating regions of elevated genetic differentiation (islands) and minimal genetic differentiation (the sea) throughout the genome.

Many questions still remain about the so-called "islands" of genetic differentiation. In particular, we know little about how big, how numerous, and how clustered these islands tend to be (Nosil and Feder, 2012). Nor do we fully understand the combinations of forces that are responsible for generating, maintaining, and degrading them. Despite the suggestion that these islands are created by selection at loci for reproductive isolating barriers or local adaptation in the presence of gene flow (Turner *et al.*, 2005), some researchers have suggested other mechanisms that might contribute to these islands, such as reduced recombination at chromosomal rearrangements (Noor and Bennet 2009). In addition, the size of regions of elevated genetic differentiation is particularly mysterious, because vastly different findings have been reported, from a few kb (Ting *et al.*, 2000) to tens of megabases (Via and West, 2008).

As of yet, few studies have been able to connect genomic patterns of heterogeneous genetic differentiation with the knowledge of the location of specific reproductive isolating loci. In this project, we will fill this gap by quantifying the size of the island of elevated genetic differentiation around a known locus for a strong prezygotic reproductive isolating barrier. This work will begin to elucidate the impact that a single strong reproductive isolating barrier can have on the differentiation of the genome in diverging species.

CHAPTER 2

The island of differentiation surrounding *pgFAR*, a strong prezygotic reproductive isolating barrier, in the European Corn Borer, *Ostrinia nubilalis*

INTRODUCTION

A major aim of current speciation research is to connect divergence observed on the phenotypic level with the divergence at the genomic level. While there are many known examples of speciation and the phenotypes contributing to reproductive isolation (Coyne and Orr, 2004), we still do not have a full understanding of how the genome diverges as two taxa split. In particular, we do not know the number, effect size or identity of loci typically responsible for reproductive isolation (Ellegren *et al.*, 2012). Nor do we understand how these loci are distributed throughout the genome or what impact they have on the divergence of the genome around them.

In the past ten years, researchers have begun to identify genes and gene regions that encode reproductive isolating barriers. These are often referred to as speciation genes, although opinions on what constitutes a "speciation gene" vary. Some consider any genes that contribute to contemporary reproductive isolation to be a speciation gene (Orr and Presgraves, 2000; Wu and Ting, 2004). Others consider speciation genes to be those that made a significant contribution to the evolution of genetic isolation between two groups, meaning that they must have evolved prior to the complete stoppage of gene flow (Nosil and Schluter, 2011). Identifying speciation genes that meet these criteria has proven difficult, particularly when employing the latter definition. In systems in which gene flow between taxa has already ended (e.g. between *Drosophila simulans* and *D. melanogaster*), it is challenging to confirm whether divergence at a particular locus contributed to speciation or developed after speciation was already complete (e.g. Brideau, *et al.* 2006; Tang and Presgraves, 2009). In other systems, the loci known to or suspected of contributing to reproductive isolation have been located within large chromosomal regions via QTL mapping, but the specific genes have yet to be identified or confirmed (Bradshaw and Schemske, 2003; Dopman *et al.,* 2005; Kronforst, *et al.* 2006; Franchini *et al.*, 2014; Wessinger *et al*., 2014; Zhang, *et al.*, 2015).

Despite these challenges, some putative speciation genes have been uncovered. Many of the earliest to be found were genes that encode intrinsic postzygotic isolation between *Drosophila* sister species. For example, two pairs of genes have been identified as responsible for hybrid inviability in *D. melanogaster* and *D. simulans* crosses: *Hybrid male rescue* (*Hmr*) and *Lethal hybrid rescue* (*Lhr*) (Barbash *et al.,* 2003; Brideau *et al.*, 2006), and *nucleoporin 96* (*Nup96*) and *nucleoporin 160* (*Nup160*) (Presgraves *et al*., 2003; Tang and Presgraves, 2009). However, because *D. melanogaster* and *D. simulans* diverged 5.4 million years ago (Tamura *et al.*, 2004) and there is no gene current gene flow between the species (Barker, 1962; Lachaise *et al.*, 1986), it is difficult to confirm that divergence at these genes preceded speciation. In contrast, the *Odysseus* (*OdsH*) gene has been identified as the cause of male hybrid sterility in the much younger *Drosophila* pair D. *simulans* and D. *mauritiana* (Perez and Wu, 1995; Ting *et al.,* 1998). Unlike the *melanogaster*/*simulans* case, these taxa diverged 250,000 years ago (Kliman *et al.*, 2000) and may not be fully reproductively isolated from each other in the field, although their predominantly allopatric distribution will limit the opportunity for gene flow. F1 hybrid females are capable of backcrossing in the laboratory, and there is evidence of gene flow that is more recent than the species split. (Nunes *et al.*, 2010). In addition to the genes reported in *Drosophila* research, several genes encoding prezygotic isolating barriers have been discovered in taxa that are thought to be experiencing contemporary gene flow. For example, Hopkins and Rausher (2011) have identified two genes involved in determining the petal color of *Phlox drummondii* (dark red) and *P. cuspidata* (light blue). Because pollinators have a consistent preference for either dark red or light blue flowers, petal color acts as a reproductive isolating barrier.

Once speciation genes are identified, an important next step is to develop an understanding how these genes influence the genome around them. That is, we should turn our attention from questions of speciation genetics to questions of speciation genomics. For species pairs that are still exchanging genes, one theory predicts that we should observe two highly similar genomes, punctuated by regions of elevated genetic difference surrounding loci encoding reproductive isolating barriers (Barton and Hewitt, 1981; Harrison, 1990; Wu, 2001). The logic is that alleles at a locus for reproductive isolation will not be able to successfully introgress into the alternate population. Any alleles that temporarily make it into the "wrong" gene pool will be quickly selected against. Nearby loci that are linked to the locus for the reproductive isolating barrier will also fail to introgress, because they are inherited as a unit with the locus under selection (hitchhiking) (Maynard Smith & Haigh, 1974). Together the locus for the reproductive isolating barrier and nearby loci will form a region of high genetic differentiation between the two species. Meanwhile, loci that are sufficiently far away from the locus for the reproductive isolating barrier will become uncoupled from that locus via recombination and are free to introgress into the alternate population, creating regions of low genetic differentiation. Nevertheless, shared ancestral polymorphism will also contribute to genomic patterns of low differentiation (Ting *et al.*, 2000; Dopman *et al.*, 2005).

Ting *et al.* (2000) found evidence in support of this phenomenon when looking at genetic differentiation in the region surrounding the *Odysseus* locus in *D. simulans*, *D. mauritiana, and D. sechellia*. In particular, they observed that there was a high degree of genetic differentiation between the three species in the immediate vicinity of *Odysseus*, but very little genetic differentiation at another marker less than 2 kb away. This result suggested that the regions of elevated genetic differentiation surrounding loci for reproductive isolating barriers may be very small.

Since then, a number of researchers have conducted genome scans, in which sequences across the entire genome of a pair of incipient species are compared in order to observe the overall pattern of genetic differentiation. In the earliest of these, Turner *et al.* (2005) compared the genomes of the M and S forms of *Anopheles gambiae* and found the anticipated pattern of regions of elevated genetic differentiation interspersed with regions of little genetic differentiation. This genome scan approach has since been used in a variety of incipient species pairs in hopes of identifying loci that are important for reproductive isolation and speciation. Scans have been conducted in organisms including flycatchers (Ellegren *et al.*, 2012), *Heliconius* (Nadeau *et al.,* 2012), sunflowers (Renaut *et al.*, 2013), whitefish (Renaut *et al.*, 2012; Gagneire *et al.*, 2013), mice (Phifer-Rixey *et al.*, 2014), sticklebacks (Hohenlohe *et al.*, 2010), and pea aphids (Via and West, 2008). Although all of the genome scan studies reveal a similar pattern of areas of genetic homogeneity punctuated by regions of elevated genetic divergence, the size of the islands varies over several orders of magnitude. Several studies have observed surprisingly large islands of genetic divergence. For example, Via and West (2008) reported regions of elevated genetic divergence on the order of 20-30 cM in host races of pea aphids and lake whitefish. Recombination rates in pea aphids are estimated at 1 cM/Mb (Wilfert *et al.*, 2007), so this is equivalent to islands of 20-30 megabases. In contrast, Ellegren *et al.* (2012) observed islands ranging from 100 kb to 3 Mb in a pair of sister flycatcher species, and Turner *et al.* (2005) found islands ranging from 37 kb to 2.1 Mb in sister mosquito races.

These regions of elevated genetic differentiation were originally dubbed “islands of speciation” and interpreted as harboring loci involved in reproductive isolation (Turner, *et al.*, 2005). However, the interpretation of these islands is more complicated than first thought (Noor and Bennet 2009;

Turner and Hahn, 2010; Dopman *et al.,* 2011; Cruickshank and Hahn, 2014). One alternate possibility is that islands of elevated genetic differentiation may contain loci for adaptation to different local environments. Loci for local adaptation may contribute to reproductive isolation if immigrants have reduced survival in the alternate habitat and/or if hybrids are unfit to live in either environment (Nosil, 2012). However, under some conditions loci for local adaptation may generate islands of differentiation even when they don't act as reproductive isolating barriers. For example, imagine two populations that were previously diverging in response to different local environments but are currently in sympatry experiencing the same environmental conditions. While the two populations experienced different environmental conditions, selection at the locus for local adaptation may have generated an island of differentiation, but now that island is being slowly degraded by uninhibited gene flow. If insufficient time has passed, a genome scan may reveal a residual island of differentiation, but this island does not represent a reproductive isolating barrier.

Another possibility is that islands of elevated genetic divergence are the result of locally reduced recombination rates around telomeres, centromeres, or chromosomal rearrangements such as inversions (Noor and Bennet, 2009). For example, because recombination is reduced or prevented within inversions (Rieseberg *et al.,* 2001), genetic differences accumulate between standard and inverted chromosomes. Consequently, for inversions that predate speciation, a region of elevated differentiation is expected even in the absence of loci for reproductive isolation or local adaptation (Noor and Bennet, 2009).

Although genome scans have enabled researchers to characterize islands of genetic differentiation, it remains a challenge to understand why such patterns are observed because of the complexities of the processes involved. An alternative approach is to attack the problem from the other direction: identify a known locus for a reproductive isolating barrier and observe the pattern of genetic differentiation that it generates. While a number of studies have looked for associations between islands of divergence and QTL underlying reproductive isolating barriers (Via and West, 2008; Hohenlohe, *et al.* 2010; Phifer-Pixey, *et al.*, 2014), few studies have looked at divergence around known speciation genes.

The European Corn Borer moth (ECB), *Ostrinia nubilalis*, is a good system in which to tackle the question of how large an island of differentiation can form around a locus for a reproductive isolating barrier. There are two strains of *O. nubilalis* that are known to hybridize (Coates *et al.*, 2013) and that are >99% reproductively isolated from each other in the field by at least seven different pre- and postzyogtic isolating barriers (Dopman *et al*., 2010). The strongest of these barriers is male orientation, which occurs in response to a pheromone blend emitted by the females (strength = 0.79) (Dopman *et al.*, 2010). E strain females produce a 99:1 blend of (E)-11-tetradecenyl acetate and (Z)-11-tetradecenyl acetate, while Z strain females produce the reciprocal blend, 3:97 E/Z. (Kochansky *et al.*, 1975; Glover *et al.*, 1987). Males of each strain preferentially fly toward and mate with females of the same strain (Linn *et al.*, 1997). Ninety nine percent of the variance in the pheromone blends of the two strains is determined by the linkage group containing the locus *pgFAR* (*pheromone gland fatty acyl reductase*), which encodes an enzyme in the pheromone biosynthetic pathway (Lassance *et al.*, 2010). Furthermore, there is no evidence of an inversion between E and Z strains in the vicinity of *pgFAR*. Thus, in ECB we have a system with a pair of naturally hybridizing species separated by a prezygotic isolating barrier with a known strength and a simple, known genetic basis, free from the influence of an inversion. These characteristics make *pgFAR* the perfect candidate in which to investigate the extent of genetic differentiation produced by the locus for a single reproductive isolating barrier in the absence of confounding variables such as inversions.

METHODS

Study System:

The European Corn Borer moth is native to Europe, North Africa, and Western Asia, and is an invasive pest in North America. The insect was first reported near Boston in 1917 and is believed to have been introduced in multiple contaminated shipments of broomcorn from Hungary or Italy between 1909 and 1914 (Caffrey and Worthley, 1927). Its range in North America currently extends from the Eastern seaboard to the Rockies and from southern Canada to the Gulf coast (Palmer *et al.*, 1985). ECB is often found on corn, but it is also known to feed on over 200 different plant species (Caffrey and Worthley, 1927). The insect is estimated to be responsible for over $1 billion dollars of damage per year in the United States (Hutchinson *et al.*, 2010).

ECB is thought to be in the early stages breaking apart into two or more species, making it a good model for speciation research. The primary division within the species is between two strains that utilize different pheromone blends in attracting mates. The females of both strains produce the same two pheromones, E- and Z-11-tetradecenyl acetate, but in different ratios: E strain females produce a 99:1 E/Z blend while Z strain females produce the reciprocal blend, 3:97 E/Z (Kochansky *et al.*, 1975; Glover *et al.*, 1987). In sustained-wind flight tunnels, E strain males preferential fly toward the pheromone blend that is characteristic of E strain females while Z strain males fly toward the Z pheromone blend (Linn *et al.*, 1997). Differential male orientation in response to pheromone blend is a powerful but incomplete reproductive isolating barrier. The strength has been found to be 0.79 in some populations (Dopman *et al.*, 2010), where a strength of 1 indicates the complete blockage of gene flow and a strength of 0 indicates completely free gene flow. When other barriers including seasonal temporal isolation, circadian temporal isolation, female discrimination, oviposition, F1 male behavioral dysfunction and F1 oviposition are taken into account, E and Z moths are 99% isolated from each other in the field (Dopman *et al.*, 2010).

The two strains are frequently found sympatrically in both Europe and North America.

Recently, the genetic basis of the difference in pheromone blend has been identified (Lassance *et al.*, 2010). A single autosomal locus, pheromone gland fatty-acyl reductase (*pgFAR*) is responsible for 99% of the difference in pheromone blends. The locus encodes an enzyme in the pheromone synthesis pathway. The two different versions of fatty-acyl reductase interact differently with precursors in the pheromone biosynthetic pathway, (E) and (Z)-11-tetradecenoyl. One version of the enzyme predominately reduces the cis isomer, ultimately resulting in a much higher proportion of the Z pheromone, while the other version of the enzyme preferentially reduces the trans isomer, resulting in a much higher proportion of the E pheromone (Lassance *et al.*, 2010).

The gene for pgFAR is 21.8 kb long and is composed of 10 exons. Exons range in size from 107 to 328 base pairs long, interspersed with introns ranging from 400 to 9,500 bases in length. The flanking genes surrounding *pgFAR* are known and their locations relative to *pgFAR* are shown in Figure 2a.

Collection of Sequencing Data:

Two different approaches were used to collect sequencing data: pooled whole genome sequencing and individual amplicon sequencing. In the

pooled approach, individuals for each population were pooled during the DNA extraction step, followed by whole genome sequencing. In the individual approach, multiplex PCR was used to generate amplicons for each individual and these were then barcoded, pooled, and sequenced via Miseq.
Both methods were used to identify single nucleotide polymorphisms (SNPs), and the allele frequencies at these SNPs were used to calculate genetic differentiation within and between the E and Z populations.

Pooled Sequencing:

Sample Collection:

Thirty-four E and 33 Z adult males were collected via pheromone trap from the Penn State Russell E. Larsen Research Center at Rock Springs, Pennsylvania (latitude 40.82◦ N, longitude 77.94◦ W, elevation of 1212). Samples were stored in an -80°C freezer until DNA extraction.

DNA Extraction, Genotyping, and Sequencing:

DNA extraction was performed individually on each moth. The moths were first frozen with liquid nitrogen, and then ground with a mortar and pestle. The DNA was extracted using the Qiagen DNeasy kit with the several modifications to the manufacturer's instructions in order to prepare high quality DNA. The samples were mixed by inverting rather than vortexting throughout the extraction process in order to avoid shearing DNA. RNase was used after the lysis step to remove any RNA contamination and each wash was performed twice. Samples were then run on a low voltage gel (40 volts), to confirm the absence of RNA in the sample. RNA containmination, if present, would have been visible as a low molecular weight schmear.

Individuals were genotyped as E or Z using a restriction enzyme assay (Coates *et al.*, 2013). E and Z pools were then created using equal quantities of DNA for each individual. DNA was sequenced on an Illumina HiSeq 2000 at Iowa State University, and 100-bp single-end reads were generated.

Bioinformatics:

The bioinformatics pipeline follows that of Kofler *et al.* (2011). FastQC was used to confirm that sequencing data were of the expected length and quality. The 3' ends of the reads were trimmed to a phred quality score of 20 using Trimmomatic v0.32, (Bolger *et al.*, 2014). The 5' ends were left untrimmed in order to facilitate the removal of duplicate reads at a later stage. All Illumina sequencing adaptors were also removed. Reads of a minimum length of 50 bases post trimming were carried forward into the alignment step.

The Burrows-Wheeler alignment tool BWA v0.6.2 (Li and Durbin, 2009) was used to align the reads to four different scaffolds, one containing *pgFAR* and three control scaffolds that did not contain any known genes that encode reproductive isolating barriers. The *pgFAR* scaffold was 627,257 bases long with *pgFAR* located from position 464,373 to 486,046. The three control scaffolds were also autosomal and were between 587 and 677 kb long.

Because the scaffolds used as references came from a Z strain individual, it was necessary to allow for a higher number of mismatches between the reads and the reference. Without this adjustment, very few E reads would align to the portions of the reference where there was a high degree of difference between the E and the Z sequence. During BWA alignment, seeding was disabled and the parameters were: -o 1 -n 8 -l 200 -e 12 -d 12.

After conversion from sam to bam using Samtools v0.1.18 (Li *et al.*, 2009), the alignments were sorted, and duplicates were removed using Picard v1.84. Samtools was used to discard reads with an alignment phred quality score below 20. The reads from the E and Z populations were compiled into an mpileup file. Because of the likelihood of spurious SNP calls surrounding indels, SNPs that occurred within five bases of an indel were masked using Popoolation2 v1201 (Kofler *et al.*, 2011b). Because areas of improbably high coverage are likely the result of read misalignment, SNPs with coverage greater than 200 were discarded. SNPs with a coverage lower than 10 were also discarded since they have little statistical power to detect genetic differentiation between the two populations. In order to avoid bias that can arise when coverage is uneven, a subsample of 10 reads were drawn from the alignment for each SNP. Popoolation2 was used to calculate F_{st} for 1 kb non-overlapping windows (step size = 1 kb).

Nucleotide diversity (π), Watterson's theta, and Tajima's D were calculated separately for E and Z populations for each of the scaffolds using the variance sliding script of Popoolation v1.2.2 (Kofler *et al.*, 2011a). Nonoverlapping 1 kb windows were again used. Tajima's D was also calculated for the E and Z pools combined.

Statistics:

Based on F_{st} results, the *pgFAR* scaffold was split into two halves: inside the island of high genetic differentiation (307 kb to 627 kb) and outside the island (1 to 307 kb). F_{st}, π, Watterson's theta, and Tajima's D values inside the island were compared to those on the *pgFAR* scaffold outside the island as well as those on the three control scaffolds using Kruskal-Wallis test followed by a Dunn's test.

F_{st} values were used to calculate the effective number of migrants per generation (Nm) (Wright 1931) using equation 1.

$$Nm = \frac{1}{4F_{st}} - \frac{1}{4} \qquad \textit{(eq. 1)}$$

All data were graphed using the ggplot2 package in R (R Core Team, 2014; Wickham, 2009).

Amplicon Sequencing:

Sample Collection and Genotyping:

Sixteen E and 20 Z moths were collected from various sites in North America and Europe (Table 1). In particular, individuals were collected from upstate New York (4 E and 10 Z), Iowa (4 Z), North Carolina (6 E and 4 Z), Italy (6 E), and Hungary (2 Z). Individuals were collected as caterpillars, pupae, and adults. The samples were genotyped as E or Z using the restriction enzyme assay described above (Coates *et al.*, 2013).

Amplicon Selection:

Amplicons were chosen based on two bacterial artificial chromosomes (BACs), each containing a portion of the *pgFAR* sequence as well as one flanking gene and some intergenic sequence. Additionally, the sequences of two flanking genes further downstream of *pgFAR* were known and used in amplicon selection. A total of 30 candidate amplicons, each between 400 and 450 bases long, were chosen. Batch Primer3 was used to identify suitable primer pairs for the amplicons within the flanking genes (You *et al.*, 2008), and Primer Blast was used for the same purpose for the other amplicons (Ye, *et al.*, 2012). Two primer pairs were chosen for each of the four flanking genes and 22 primer pairs were chosen such that the remaining amplicons would be roughly evenly spread throughout the two BACs. Standard PCR was used to identify which of the 30 primer pairs were able to successfully amplify DNA and thus should be carried forward into the multiplex PCR. All eight of the flanking gene primer pairs were carried into the multiplex step as well as nine of the remaining primer pairs.

Multiplex PCR:

Primers were designed which contained the normal primer sequence along with an overhang sequence that would act as an adaptor for the subsequent addition of barcodes. A different overhang sequence was used for the forward primer and the reverse primer, such that a different barcode could be attached to each end of the amplicon. Multiplex PCR was used to amplify all 17 loci at the same time for each individual moth (Qiagen Multiplex PCR Kit).

Library Preparation and Sequencing:

A second round of PCR was used to add barcodes to each amplicon indicating which individual it had been derived from. A second set of primers were designed which were composed of 1) either the forward or reverse overhang sequence used in the first round of PCR, 2) a barcode, and 3) an adaptor which allowed the amplicon to be compatible with the Illumina flow cell. There were 8 different barcodes used for the forward primers and another 6 barcodes used for the reverse primers, allowing for 48 different unique combinations of one forward and one reverse barcodes. Each individual was given one of these unique combinations such that the identity of amplicons could subsequently be determined post sequencing.

Following PCR, ampure beads were used to purify DNA. The DNA samples were then sent to the Tufts University Core Facility where they were tested for quality control, pooled, and sequenced. Paired-end sequencing was conducted on a Miseq machine to generate 250 bp reads.

Bioinformatics:

The bioinformatics pipeline followed De Wit *et al.* (2012). FastQC was used to confirm that sequencing data were of the expected length and quality. Both the 3' and 5' ends of the reads were trimmed to a phred quality score of 20 using Trimmomatic v0.32 (Bolger *et al.*, 2014). All Illumina sequencing adaptors were also removed in the trimming step. BWA v0.6.2 (Li and Durbin, 2009) was used to align the reads to the *pgFAR* scaffold (which was the same as that used in the pooled sequencing portion of the project above). The parameters used were -o 1 -n 0.01 -l 30 -e 12 -d 12 -t 2. Samtools v0.1.18 (Li *et al*., 2009) was used to discard reads with an alignment phred quality score below 20.

Data from all E individuals were merged together as was data from all Z individuals. IndelRealigner tool from the Genome Analysis Toolkit was used to do local realignment of reads near indels in order to correct for misalignments that can occur these areas (McKenna *et al.*, 2010; DePristo *et al.*, 2011). SNPs were identified using the GATK UnifiedGenotyper. Then the GATK VariantFiltration tool was used to filter out SNPs that were are likely to be artifacts and preserve high quality SNPs (De Wit, *et al.* 2012, Van der Auwera *et al.*, 2013). Finally, F_{st} values were calculated using the HierFstat package in Rstudio (Goudet, 2005).

RESULTS

Pooled Sequencing:

E and Z individuals from sympatric locations were pooled and sequenced. The mean per base phred quality score ranged from 33-39, depending on position within the read. Following the removal of duplicates, there were 156,281 (E) and 161,263 (Z) high quality alignments to the *pgFAR* scaffold, resulting in a median coverage of 16 (E) and 18 (Z) (Table 2) A similar number of high quality alignments and similar median coverages were observed in the control scaffolds (Table 2). A total of 26,657 single nucleotide polymorphisms (SNPs) were identified in the *pgFAR* scaffold, corresponding to approximately 4.2% of bases. A similar proportion of bases contained SNPs in the control scaffolds (Table 2).

F_{st} was calculated throughout all four scaffolds for non-overlapping 1 kb windows. A 320 kb long region of elevated F_{st} was observed to be roughly centered on the *pgFAR* locus (Figure 2a). This observation is in marked contrast with the control scaffolds, in which F_{st} was consistently low at all windows throughout the scaffold (Figure 2b-d). At the resolution of 1 kb windows, the region of elevated F_{st} appears to be composed of one primary peak centered at 505 kb, which is roughly 160 kb in length and two smaller secondary peaks centered at 325 kb and 608 kb, which are 50 kb and 90 kb in length, respectively. The ten coding regions of *pgFAR* fall between 464 and 486 kb on the scaffold. A number of loci surrounding *pgFAR* are known

(Figure 2a), including *lipase 3* (460 to 462 kb), *vitellogenin* (514 to 523 kb), *mitochondrial aldehyde dehydrogenase* (580 to 588 kb), *lipase* (588 to 590 kb), *mitochondrial sodium/hydrogen exchanger* (607 to 609 kb), an unnamed gene (611 to 613 kb), a gene for Sfmbt-like polycomb protein (617 to 624 kb).

F_{st} varies considerably within the region of elevated differentiation. Windows with low and moderate average F_{st} values are interspersed with windows of much higher F_{st} values. Based on F_{st} data, the *pgFAR* scaffold was split into two halves: inside the island of high genetic differentiation (307 kb to 627 kb) and outside the island (1 to 307 kb). F_{st} inside the island is significantly greater than outside the island (median: 0.109 vs. 0.056, respectively) as well as significantly greater than the three control scaffolds (medians: control scaffold 1: 0.056; control scaffold 2: 0.056; and control scaffold 3: 0.058; Kruskal-Wallis, chi-squared = 285.3439, df = 4, p-value < 2.2×10^{-16}, followed by Dunn's test, p value < 0.0001 for all comparisons with the inside island group, Figure 3).

Migration rates were estimated based on F_{st} data. For the control scaffolds and the region outside the island of differentiation on the *pgFAR* scaffold, the median migration rate was found to be 4.15 individuals per generation (Figure 4). In contrast, the effective migration rate within the island on the *pgFAR* scaffold is only 2.04 individuals per generation.

Nucleotide diversity (π) was calculated within E and Z populations for each of the four scaffolds (Figure 5). In the E population, median nucleotide diversity inside the *pgFAR* island was significantly lower than that of the other four regions (Kruskal-Wallis, inside island median = 0.014, median for other regions range from 0.016 to 0.019, chi-squared = 68.0606, df = 4, p-value = 5.825 x 10^{-14}, followed by Dunn's test: p value < 0.0004 for all comparisons with the inside island group, Figure 5a). A similar pattern was observed in the Z population; nucleotide diversity was found to be significantly lower inside the island than in the other four regions Kruskal-Wallis, inside island mean = 0.012, median for other regions range from 0.016 to 0.019, chi-squared = 98.9175, d.f. = 4, p-value < 2.2 x 10^{-16}, followed by Dunn's test: p value < 0.0001 for all comparisons with the inside island group, Figure 5b.)

Results for Watterson's estimator of polymorphism was similar to that calculated by π (Figure 6). In the E population, median Watterson's theta was significantly lower than that of the other four regions (Kruskal-Wallis, inside island median: 0.011, median for other regions range from 0.016 to 0.020, chi-squared = 124.1409, df = 4, p-value < 2.2 x 10^{-16}, followed by Dunn's test: p value < 0.0001 for all comparisons with the inside island group, Figure 6a). The same pattern was observed in the Z population; the median value of Watterson's theta was significantly lower inside the island compared to the other regions (Kruskal-Wallis, inside island median: 0.011, median for other regions ranges from 0.016 to 0.020, chi-squared = 164.3073, df = 4, p-value < 2.2 x 10^{-16}, followed by Dunn's test: p value < 0.0001 for all comparisons with the inside island group, Figure 6b).

The two measures of polymorphism were used to calculate Tajima's D for E and Z populations (Figure 7). The median Tajima's D values within the island, outside of the island on the same scaffold, and on control scaffolds were all between -0.043 and -0.120 for both E and Z populations. There was no significant difference between any of the regions for either the E or the Z populations (E: Kruskal-Wallis chi-squared = 8.6403, d.f. = 4, p-value = 0.07075, Z: Kruskal-Wallis chi-squared = 8.506, d.f. = 4, p-value = 0.07471, Figure 7).

We also measured Tajima's D when the E and Z populations were combined and viewed as a single population (Figure 8). For the combined populations, we found that Tajima's D was significantly more positive (i.e. closer to zero) inside the island (D = -0.233) compared to the other regions outside of the island (D ranged from -0.316 to -0.409, Kruskal-Wallis test, chi-squared = 30.8451, df = 4, p-value = 3.292 x 10^{-6}, followed by Dunn's test: p value < 0.0069 for all comparisons with the inside island group, Figure 8).

Amplicon Sequencing:

Seventeen loci in the *pgFAR* region were amplified from E and Z individuals and sequenced. Of these, reads from nine of the loci were successfully aligned to the *pgFAR* scaffold. The mean per base phred quality score ranged from 32-38, depending on position within the read. The mean number of high quality reads that aligned to the scaffold was 77,383 for E individuals and 98,551 for Z individuals (Table 3). The amplicons were 400-

450 bases long, and there were an average of 8.22 SNPs was observed per amplicon (Table 3), making the SNP rate approximately 2%.

F_{st} was calculated for each SNP identified in each amplicon. Although some particular SNPs had low F_{st} values, in general F_{st} was elevated throughout the region (Figure 9), matching what was observed in the pooled data (Figure 2a). Average F_{st} was 0.18, which is roughly on par with what was observed in the pooled data (Figure 2a). The location of the F_{st} peaks in the amplicon data matches that of the pooled data, with peaks around 505 and 608 kb (There were no amplicons that overlapped with the 325 kb peak from the pooled data.) The peak around 608 kb corresponds to a gene that has been identified as a mitochondrial sodium/hydrogen exchanger. The locations of other genes relative to *pgFAR* are depicted in Figure 9.

DISCUSSION

Understanding speciation is one of the major goals of evolutionary biology. Genome scans in dozens of hybridizing species pairs have repeatedly revealed a pattern of heterogeneous genetic differentiation throughout the genome (e.g. Turner *et al.,* 2005; Via and West, 2008; Renaut *et al*., 2012; Gagneire *et al.*, 2013; Martin *et al.,* 2013; Nadeau *et al.*, 2013; Renaut *et al.*, 2013; Phifer-Rixey *et al.*, 2014). Some researchers have hypothesized that loci for reproductive isolating barriers underlie these islands of high genetic differentiation (Turner *et al.,* 2005), but others argue that alternative forces may be responsible (Noor and Bennet 2009; Turner and Hahn, 2010;

Dopman *et al.,* 2011; Cruickshank and Hahn, 2014). Here we take the opposite approach. We start with a known evolutionary process and observe the pattern of genetic differentiation it generates in the genome. In the European Corn Borer system we have a pair of naturally hybridizing species, separated by a reproductive isolating barrier of a known strength (Dopman *et al*., 2010) and a known genetic basis (Lassance *et al.,* 2010). In this study, we observe the extent of region of elevated genetic differentiation that is generated around the locus for this barrier, *pgFAR*.

In the F_{st} analysis, we observed an island of approximately 320 kb surrounding the *pgFAR* locus with a maximum value of 0.67 (Figure 2a). In order to investigate the role of selection in creating this island, we examined Tajima's D (Tajima, 1989) across the four scaffolds. Tajima's D equals zero under neutrality and is based on the difference between nucleotide diversity (π), a measure of the total diversity within a population, and Watterson's theta (θ_W), an adjusted measure of the number of polymorphic sites (Watterson, 1975). In both the E and Z populations, we observed a Tajima's D value within the *pgFAR* island that is close to zero and not significantly different from the Tajima's D values outside of the *pgFAR* island (Figure 7). Additionally, we observed that both π and θ_W were significantly lower inside the *pgFAR* island when compared to the other scaffolds (Figures 5 and 6). Together these patterns suggest two important points. First, the arrival of the pheromone blend reproductive isolating barrier led to selective sweeps inside the *pgFAR* island, which reduced diversity as measured by both π and

θ_W within the strains. Second, these sweeps happened sufficiently long ago that the relationship between π and θ_W has corrected itself, resulting in a neutral value for Tajima's D. If the selective sweeps had occurred more recently, we would have expected to see a negative Tajima's D within the *pgFAR* island. This would occur because the recent selective sweep would have wiped out much of the ancestral variation and new mutations that arose since the sweep would still exist at low frequencies, shifting the allele frequency spectrum towards higher values of θ_W compared to π. We also examined Tajima's D when the E and Z populations were combined and viewed in a single population (Figure 8). In this case, the average Tajima's D value within the *pgFAR* island is negative, but it is significantly more positive (i.e. closer to 0) than the average Tajima's D values for the comparison regions. Such a locus-specific effect is best explained by reduced migration in the *pgFAR* interval (Nm = 2.04) and higher migration elsewhere (Nm: 3.83 to 4.20) (Figure 4).

The island size at *pgFAR* is roughly similar to average island size observed in species pairs of flycatchers (Ellegren *et al.*, 2012), butterflies (Nadeau *et al.*, 2012), and migratory songbirds (Ruegg *et al.*, 2014). However, it is several orders of magnitude smaller than the islands reported around QTL for reproductive isolating barriers in pea aphids (Via and West, 2008) and whitefish (Renaut, 2012) and several order of magnitude larger than the region of elevated differentiation observed around the *Odysseus* locus in *D. simulans* and *D. mauritiana* (Ting *et al*. 2000). What can account for this wide range of sizes in islands of genetic divergence? For islands that arise around a reproductive isolating barrier, at least three relevant factors are thought to control the size of the island: strength of reproductive isolation at the locus, recombination rate, and time since the isolating barrier arose. The stronger the selection against the "wrong" allele, the more quickly the haplotypes that contain that allele will be removed and thus not available for recombination, maintaining a larger region of elevated differentiation. In contrast, a high recombination rate reduces the size of an island because it separates neutral alleles from hitchhiking with the alleles under selection. The greater the recombination rate, the faster this is expected to happen and the smaller the island size. Likewise, the more time that has passed since the reproductive isolating barrier arose, the greater the likelihood that recombination has separated any particular neutral loci from the locus under selection, and thus the smaller the island.

Assuming that patterns found at and near *OdsH* by Ting *et al.* (2000) were representative of patterns on the rest of the chromosome, we can speculate about how these factors might help explain discrepancy in the size of the islands surrounding *pgFAR* and *Odysseus*. The locus for *pgFAR* contributes to a reproductive isolating barrier that blocks 79% of gene flow (Dopman, *et al.* 2010) while *OdsH* blocks 25% of gene flow (Nosil and Schluter, 2011). The 3-fold greater strength *pgFAR* as a reproductive isolating barrier means that there would be less introgression possible at this site, and thus a larger island would be expected, all other things being equal.

The E and Z strains of ECB are estimated to have split between 75,000 and 150,000 years ago (Malausa *et al.,* 2007); at two generations per year, this corresponds to approximately 150,000 to 300,000 generations. In contrast, *D. simulans* and *D. mauritiana* are estimated to have diverged approximately 260,000 years ago (Kliman, *et al.* 2000), corresponding to roughly 2.6 million generations. If we assume that both *pgFAR* and *Odysseus* diverged early in the speciation process, there has been considerably more time (in terms of number of generations) for the island of divergence around *Odysseus* to be whittled away. The average recombination rate in ECB has been estimated to be 3.45 cM/Mb (Dopman *et al*., 2004). The average recombination rate for *D. mauritiana* is similar at 3.1 cM/Mb (Wilfert *et al.,* 2007) and that of *D. simulans* is slightly lower (True *et al.,* 1996). In total, two of these observations are consistent with the smaller island observed at *OdsH*: 1) the smaller strength of reproductive isolation, and 2) more time, in terms of generations, for the island of differentiation to be reduced through recombination. The third observation, only slightly lower recombination rates in the *Drosophila* species, is not fully consistent with the different island sizes observed in the two taxa.

The *Heliconius* radiation offers a second instructive comparison. *Heliconius* use wing patterns as the basis for mate recognition, so loci for wing coloration contribute to reproductive isolation (Jiggins *et al.*, 2001). The location of several clusters of genes which contribute to wing coloration have been identified: *HmYb*, *HmSb*, and *HmN,* which contribute to yellow and white elements (Jiggins *et al.,* 2005; Ferguson *et al.*, 2010) and *HmB* and *HmD*, which contribute to red and orange elements (Baxter *et al.*, 2008). Nadeau *et al.*, (2012) examined the degree of differentiation between two Peruvian *H. melpomene* races, *H. m. aglaope* and *H. m. amaryllis*, around these two clusters of genes. They observed islands of differentiation on the order of 500 kb, similar to what we found surrounding *pgFAR*. Recombination rate in other *H. melpomene* races has been observed to be 5.55 cM/Mb (Jiggins *et al*., 2005), slightly greater than that of ECB. Reproductive isolation between the two races is not strong; *H. m. amaryllis* males have a mild preference for *H. m. amaryllis* females, while *H. m. aglaope* males have no preference (Merrill *et al.*, 2011). Based on these figures, we'd expect the islands surrounding the wing color loci to be moderately smaller than that around *pgFAR*; instead, they are moderately larger. One possible explanation is that perhaps not enough time has passed for the work of recombination to fully chip away at the island of differentiation within the *H. melpomene* races.

A fourth factor that may influence island size is the presence of more than one target of selection in the same stretch of chromosome. Divergence hitchhiking theory proposes a mechanism by which islands of differentiation could grow over time, rather than shrink in the face of recombination, when there are multiple targets of selection near each other (Via, 2012). The theory asserts that selection at a locus for a reproductive isolating barrier will create a region of reduced effective migration around the locus via hitchhiking. Once effective migration rates are reduced, loci within this region that experience only weakly divergent selection may now become differentiated, despite the homogenizing effects of ongoing gene flow. These second locus will, in turn, influence the effective migration rate experienced by loci further downstream and the region of differentiation can spread along the genome. Via argues that without divergence hitchhiking, we should expect islands of differentiation around loci for reproductive isolating barriers to be only a few kilobases long (Via, 2012) such as those observed surrounding *Odysseus*. If this is true, then divergence hitchhiking may explain the discrepancy between the size of the regions of differentiation observed around *pgFAR* and the *Heliconius* wing color loci vs. *Odysseus*.

If divergence hitchhiking is operating, there may be signatures in the structure of islands of divergence. At the resolution created by using 1 kb windows, the island surrounding *pgFAR* appears to be composed of one primary peak with a length of roughly 160 kb and two secondary peaks on either side of approximately 50-90 kb (Figure 2a). The multiple peak pattern was less pronounced when window size was increased (e.g., to 10 kb). Nevertheless, the presence of multiple peaks of differentiation could be consistent with divergence hitchhiking. In a divergence hitchhiking scenario, selection at *pgFAR* would create a region of reduced effective migration around it. This region of reduced migration would allow for genetic differences to accumulate at nearby loci with weakly divergent selection, potentially creating the observed secondary peaks in F_{st} (Figure 2a). The low points in between the peaks would be caused by loci that are not under differential selection and are too far from either *pgFAR* or the secondary peak loci to experience a hitchhiking effect. Proponents argue that divergence hitchhiking could ultimately result in islands of differentiation that are tens of megabases long, such as those observed in pea aphids (Via and West, 2008) and whitefish (Renaut, 2012), however, the island surrounding *pgFAR* is much smaller than this. If divergence hitchhiking is happening in ECB, the E and Z strain moths must be in a very early stage relative to pea aphids and whitefish.

Another explanation for the multiple peaked patterns is that regions of low differentiation within islands correspond to a high degree of ancestral polymorphism that is shared between the diverged taxa (Via and West, 2008). Under this scenario, even though *pgFAR*-Z haplotypes are quickly removed from E gene pool and vice versa, both E and Z haplotypes have similar allele frequencies at SNPs in these regions. Thus, even as the "wrong" haplotypes are continuously purged, genetic differentiation in these regions remains low. A third possibility is that pheromone blend is controlled more than one gene. Lassance *et al.* report that 99% of the variance in pheromone ratios is controlled by linkage group 31, but the possibility that more than one gene is involved remains (2010). In this scenario, each locus would produce an island of 50-150 kb region of genetic differentiation and homogenization is allowed to occur in between.

The flanking loci surrounding *pgFAR* may provide some hint as to which process or combination of processes are responsible for the multi peak pattern of differentiation around *pgFAR*. While the F_{st} peak at position 325 kb does not correspond with any known gene (Figure 2a), the peak at 608 kb corresponds to the locus for the *mitochondrial sodium/hydrogen exchanger* (*NHA2*) (Figure 9). The protein encoded by this gene is responsible for maintaining the balance of ions inside mitochondria and thus controlling the pH and volume of the organelles (Fuster *et al.,* 2008). Such a gene is presumably critical for the healthy functioning of the cell, and thus could contribute to reproductive isolation through a Dobzhansky-Muller incompatibility. If differences were to accumulate in this locus in the two ECB lineages, hybrids might be inviable if the allele from each lineage did not function properly in the opposing genetic background. The same argument applies to several other loci in the vicinity of *pgFAR*, such as *vitellogenin,* located from 514-523 kb. The *vitellogenin* locus is essential for the formation of yolk in eggs. Dopman *et al.* (2010) observed that ECB Z females mated with E males laid significantly fewer eggs when compared to the within strain cross. This finding is consistent with a Dobzhansky-Muller incompatibility that reduces the success of hybrids when vitellogenin from one strain is dysfunctional in the presence of the alternative genetic background. Similar logic applies to the locus that encodes a Sfmbt-like polycomb protein, located at 617 to 624 kb, a likely transcription regulator involved in development (Wu *et al.*, 2007).

In addition, there are loci for two different lipases in the vicinity of *pgFAR*. Lipases are known to play a role in the pheromone biosynthetic pathway in the silkworm *Bombyx mori*, a close relative of ECB, where they split triglycerides into pheromone precursors (Du, *et al.* 2012). It is reasonable to assume that lipases play a similar role in the ECB. Lipase is involved in an early stage of the pheromone biosynthesis pathway, before the two distinct isomers of (E/Z)-11-tetradecenyl acetate are formed. However, it is possible that different E and Z lipase alleles could contribute to reproductive isolation by determining the total amount of pheromone produced.

Despite the substantial amount of work that has been devoted to 1) identifying genes that contribute to reproductive isolation and 2) identifying genomic regions of elevated genetic divergence, the relationship between the two remains murky. Some studies have observed islands of elevated divergence that correspond with previously known QTL for reproductive isolating barriers (e.g. Via and West, 2008; Hohenlohe *et al.,* 2010; Phifer-Rixey *et al.,* 2014). However, other empirical work has suggested that islands of divergence do not always contain reproductive isolating barriers and that reproductive isolating barriers do not always generate islands that are detectable in a genome scans. For example, in the M and S forms of *Anopheles gambiae*, a recent introgression of an insecticide resistance mutation residing within a large island of divergence has caused the island to become completely homogenized (Clarkson *et al.,* 2014). Despite the disappearance of the island of divergence, reproductive isolation was not reduced, leading the authors to infer that the island did not contain a locus that was important for reproductive isolation. Conversely, in a recent study, researchers compared the results of genome scan of a pair of migratory songbird species with the locations of loci linked to migratory traits involved in reproductive isolation. They found that three out of four loci linked to migration related reproductive isolating barriers were located outside of the islands of elevated divergence (Ruegg *et al.,* 2014). Presumably, there must be at least small regions of genetic divergence at the loci that encode the different migratory behaviors, but the regions of genetic differentiation generated around these loci must too small to be detected by genome scans. Clearly, more work needs to be done to determine why some speciation genes generate kilobase or megabase islands while others do not.

Our study contributes to the debate by examining the size of the island of genetic differentiation around a known speciation gene in a system that is experiencing ongoing gene flow. We observed a 320 kb island surrounding *pgFAR* with a maximum F_{st} of 0.67. This work suggests that a strong, relatively recent prezygotic reproductive isolating barrier should be expected to generate a region of differentiation of several hundred kb, in the absence of other contributing factors such as a chromosomal inversion. Islands that are much larger or smaller may indicate the presence of other mechanisms or be the result of more ancient reproductive isolating barriers. As future work in this system uncovers more regions of genetic differentiation, those regions of a few hundred kilobases will be reasonable places to look for reproductive isolating barriers. In contrast, if future work uncovers islands of divergence that are megabases long, we should inquire about the presence of another mechanism such as an inversion or extensive divergence hitchhiking that would lead to the formation of an island of this size.

A big question yet to be answered is the field of speciation is how much genetic differentiation is generated for a given amount of reproductive isolation, particularly in a speciation-with-gene-flow type situation? Is there a linear or nonlinear relationship between the strength of the reproductive isolating barrier and the size of the island of differentiation that is generated? Perhaps a small amount of reproductive isolation is sufficient to generate a large island of genetic differentiation with limited impact of any additional reproductive isolation. Or, conversely, perhaps an island of differentiation of appreciable size is only generated under conditions of very strong reproductive isolation (Figure 10). Ultimately, we will be able to determine this relationship when we have results from a variety of systems with reproductive isolating barriers of variety of strengths. Together, such studies will yield insight into how genetic differences accumulate as species diverge.

TABLES

Table 2.1: Identity of individuals used in amplicon study.

E		Z	
Geneva, NY	4	Geneva, NY	7
Weeksville, NC	6	Eden, NY	3
Piacenza, Italy	6	Ames, IA	4
		Fletcher, NC	4
		Kety, Hungary	2
Total	16		20

Table 2.2: Summary statistics for Rock Springs reads aligned each of the scaffolds. The E pool was created from 34 individuals and contained 143,710,538 raw reads. The Z pool was created from 33 individuals and contains 157,746,984 raw reads.

	pgFAR scaffold		Control scaffold 1		Control scaffold 2		Control scaffold 3	
	E	Z	E	Z	E	Z	E	Z
Length of scaffold (bp)	627,257		644,832		676,544		586,973	
High quality reads that aligned to a single place in the scaffold	2,595,235	2,787,226	2,513,575	2,633,891	1,829,374	1,956,872	2,583,733	2,724,671
Reads remaining after duplicates have been removed and which have a high quality alignment score	156,281	161,263	152,022	148,610	135,992	133,776	157,080	155,490
Mean coverage	30	31	28	28	24	24	32	32
Median coverage	16	18	16	15	14	13	19	18
Number of SNPs found	26,657		28,526		27,563		27,642	
Percentage of bases that contain a SNP	4.2%		4.4%		4.1%		4.7%	

Table 2.3: Summary statistics for amplicon reads aligned to the *pgFAR* scaffold.

	E	Z
Sample size	16	20
Mean number of raw reads per individual	241,677	266,899
Mean number of high quality reads that aligned to the scaffold per individual	77,383	98,551
Mean number of high confidence SNPs per amplicon	8.22	

FIGURES

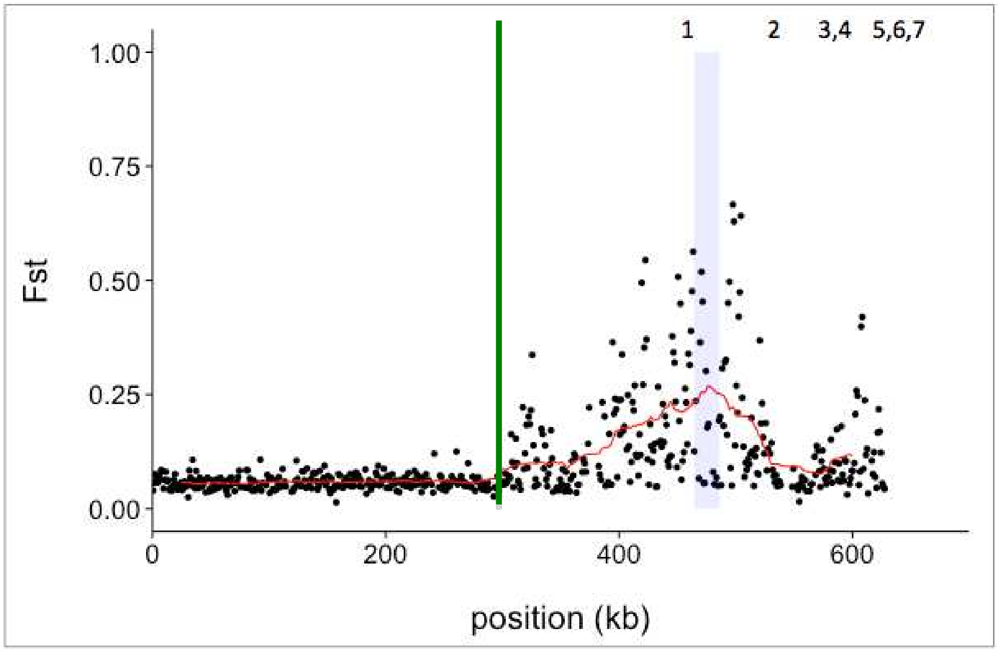
Figure 2.1: Picture of female (left) and male (right) adult ECB (a). Picture of ECB larva that has damaged a corn stalk (b). Sources: a) http://www.ent.iastate.edu/pest/cornborer/female-and-male-adult-moths and b) http://www.aaas.org/sites/default/files/migrate/uploads/1007sp_gm_corn.jpg

a)

b)

Figure 2.2: F_{st} values for the Rock Springs population. Each point represents the mean F_{st} for all SNPs within a 1kb window. The red line represents a rolling mean of 50 windows. a) *pgFAR* scaffold. The shaded blue bar represents the location of *pgFAR*. The numbers indicate the location of flanking genes: 1, lipase 3; 2, vitellogenin; 3, mitochondrial aldehyde dehydrogenase; 4, lipase; 5, mitochondrial sodium/hydrogen exchanger; 6, unnamed cds; 7, Sfmbt-like polycomb protein. b-d) control scaffolds 200, 422, and 437 respectively. The green line indicates the split between the region considered to be inside the *pgFAR* island (right) and the region considered to be outside the *pgFAR* island (left).

a)



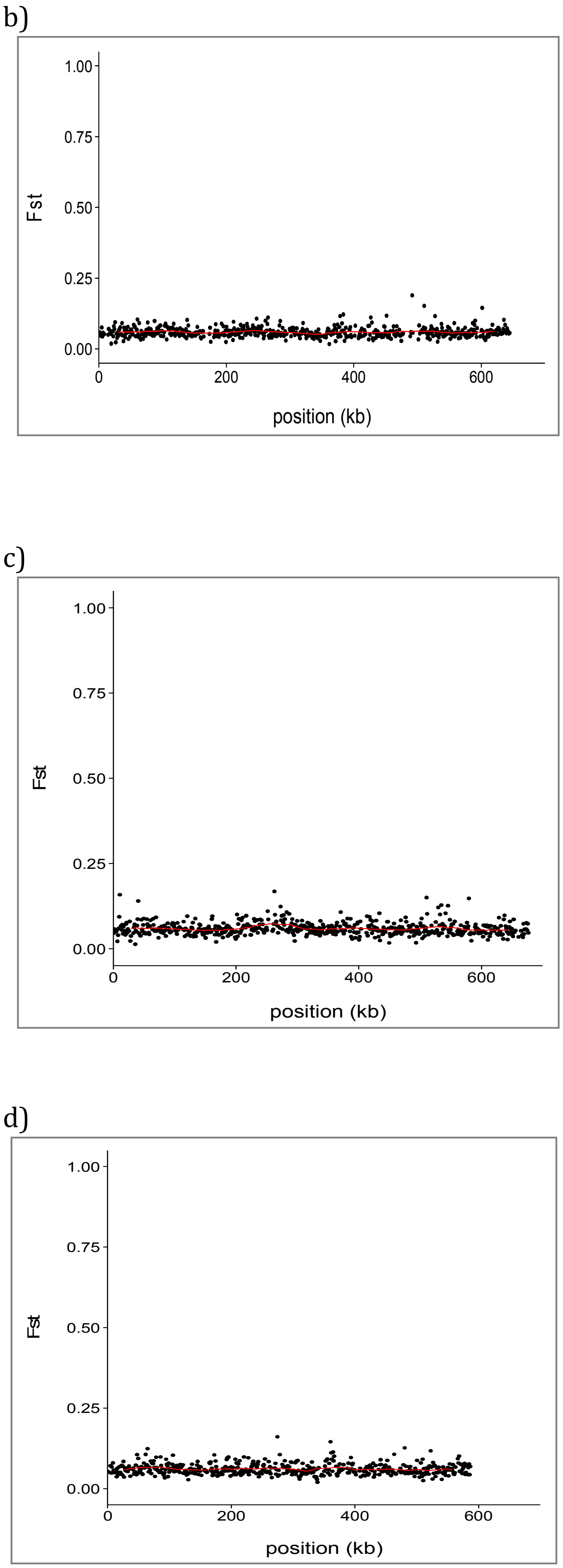
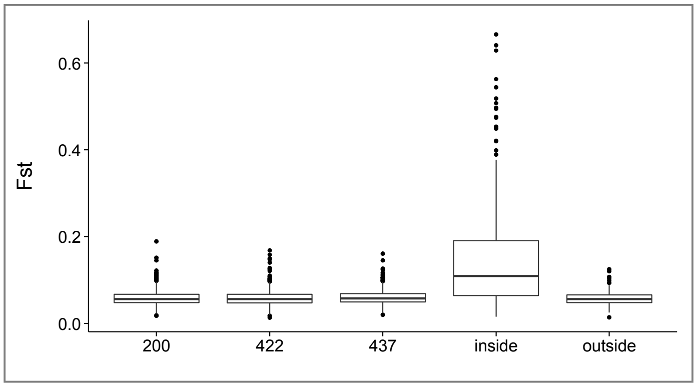

b)
1.00
0.75
0.50
0.25
0.00
Fst
0
200
400
600
position (kb)
c)
1.00
0.75
0.50
0.25
0.00
Fst
0
200
400
600
position (kb)
d)
1.00
0.75
0.50
0.25
0.00
Fst
0
200
400
600
position (kb)

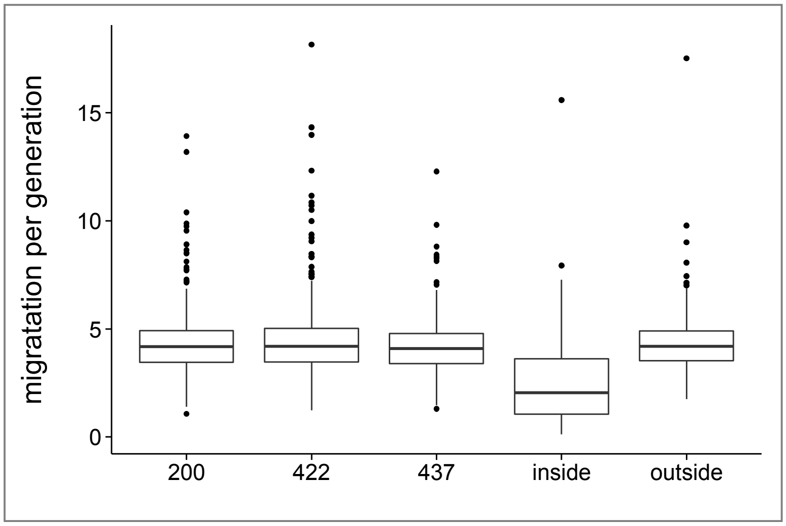
Figure 2.3: Boxplot of the F_{st} values for the Rock Springs population. Boxplots were constructed for the data points inside the *pgFAR* island, outside the island on the *pgFAR* scaffold, and for each of the three control scaffolds. P-values for the five-way comparison were generated with a Kruskal-Wallis test, followed by a Dunn's test.



p-values:

	200	422	437	inside
422	0.2752	-	-	-
437	0.0853	0.0246	-	-
Inside	0.0000	0.0000	0.0000	-
Outside	0.3180	0.4961	0.5557	0.0000

Figure 2.4: Effective migration in Rock Springs population. Effective number of migrants in each window was calculated based on the mean F_{st} value in that window using equation (1). Boxplots were constructed for the data points inside the *pgFAR* island, outside the island on the *pgFAR* scaffold, and for each of the three control scaffolds. P-values for the five-way comparison were generated with a Kruskal-Wallis test, followed by a Dunn's test.

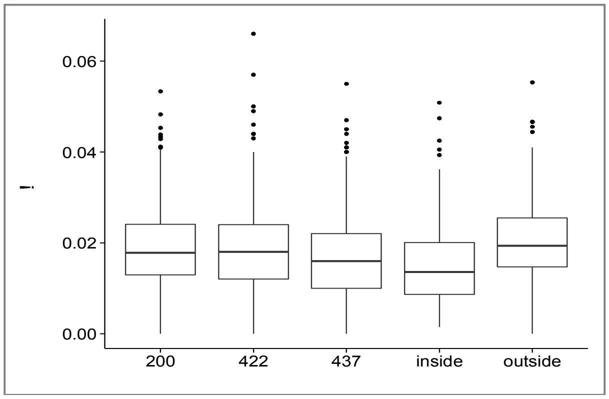


p-values:

	200	422	437	inside
422	0.2752	-	-	-
437	0.0853	0.0246	-	-
Inside	0.0000	0.0000	0.0000	-
Outside	0.3180	0.4961	0.0557	0.0000

Figure 2.5: Nucleotide diversity (π) for Rock Springs populations. Boxplots were constructed for the data points inside the *pgFAR* island, outside the island on the *pgFAR* scaffold, and for each of the three control scaffolds, a) E population and b) Z population. P-values for the five-way comparison were generated with a Kruskal-Wallis test, followed by a Dunn's test.

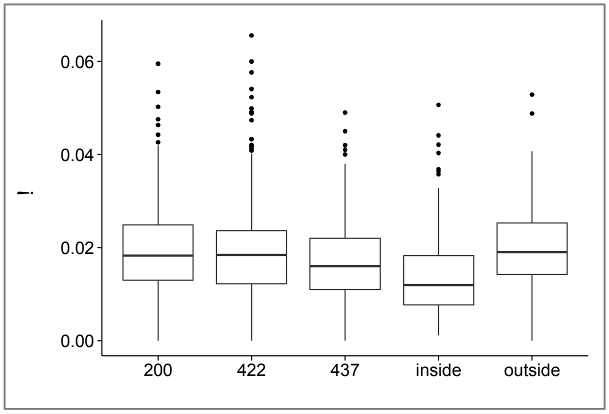
a)



p-values:

	200	422	437	inside
422	0.1348	-	-	-
437	0.0004	0.0104	-	-
Inside	0.0000	0.0000	0.0004	-
Outside	0.0083	0.0005	0.0000	0.0000

b)

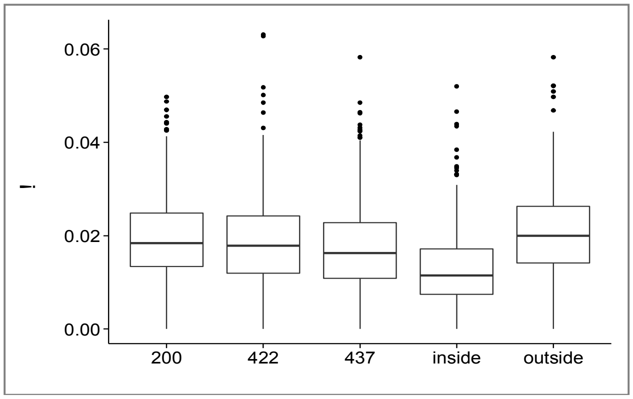


p-values:

	200	422	437	inside
422	0.2150	-	-	-
437	0.0003	0.0041	-	-
Inside	0.0000	0.0000	0.0000	-
Outside	0.0970	0.0197	0.0000	0.0000

Figure 2.6: Watterson's theta (θ) for Rock Springs populations. Boxplots were constructed for the data points inside the *pgFAR* island, outside the island on the *pgFAR* scaffold, and for each of the three control scaffolds, a) E population and b) Z population. P-values for the five-way comparison were generated with a Kruskal-Wallis test, followed by a Dunn's test.

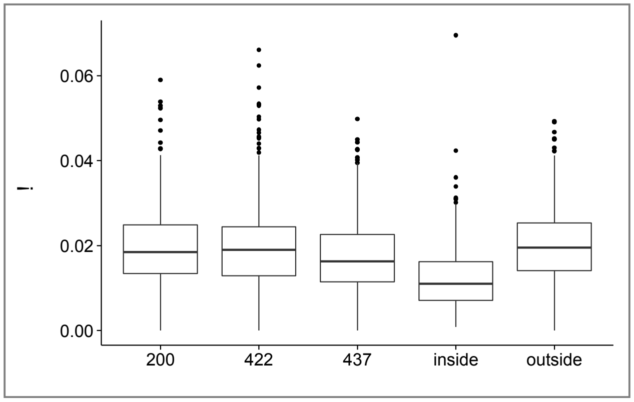
a)



p-values:

	200	422	437	Inside
422	0.1293	-	-	-
437	0.0002	0.0074	-	-
Inside	0.0000	0.0000	0.0000	-
Outside	0.0247	0.0021	0.0000	0.0000

b)

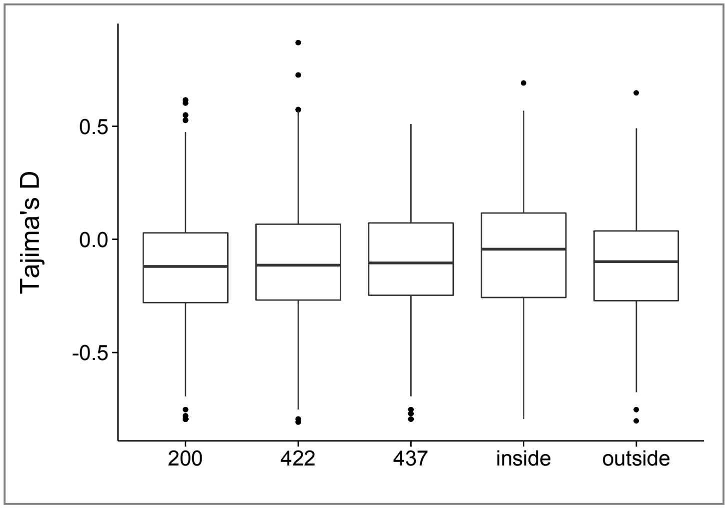


p-values:

	200	422	437	Inside
422	0.3910	-	-	-
437	0.0005	0.0012	-	-
Inside	0.0000	0.0000	0.0000	-
Outside	0.1053	0.0699	0.0000	0.0000

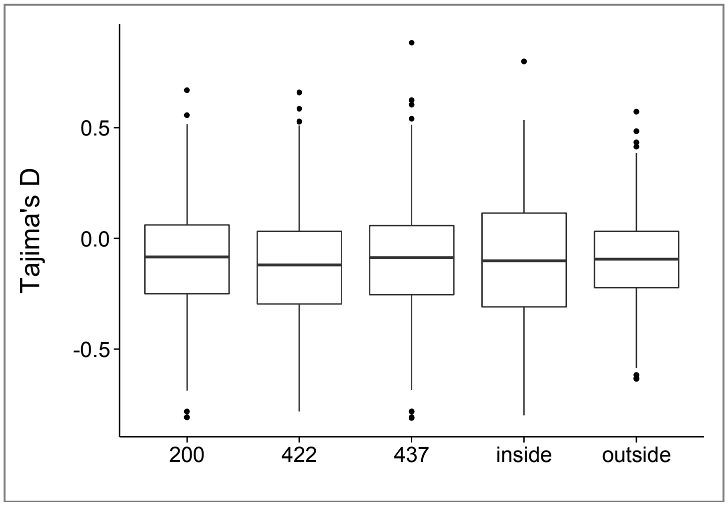
Figure 2.7: Tajima's D for the Rock Springs populations. Boxplots were constructed for the data points inside the *pgFAR* island, outside the island on the *pgFAR* scaffold, and for each of the three control scaffolds, a) E population and b) Z population. A Kruskal-Wallis test was used to test for significant differences.

a)



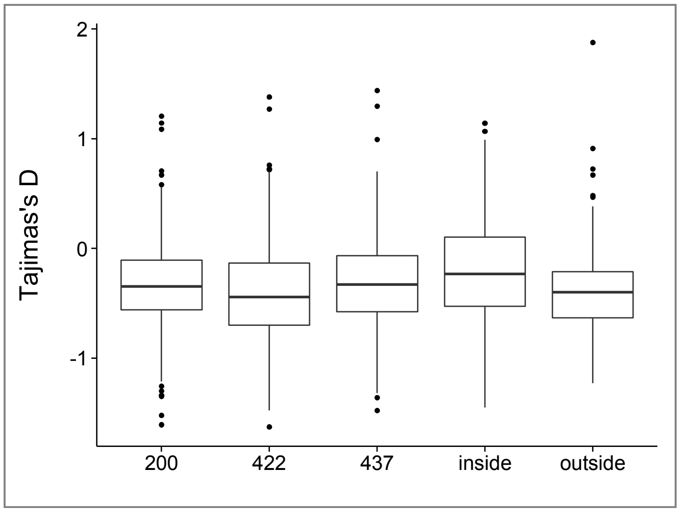
Kruskal-Wallis chi-squared = 8.6403, df = 4, p-value = 0.07075

b)



Kruskal-Wallis chi-squared = 8.506, df = 4, p-value = 0.07471

Figure 2.8: Tajima's D for the Rock Springs population with E and Z individuals combined. Boxplots were constructed for the data points inside the *pgFAR* island, outside the island on the *pgFAR* scaffold, and for each of the three control scaffolds. P-values for the five-way comparison were generated with a Kruskal-Wallis test, followed by a Dunn's test.



p-values:

	200	422	437	Inside
422	0.0006	-	-	-
437	0.3468	0.0002	-	-
Inside	0.0024	0.0000	0.0069	-
Outside	0.0048	0.4808	0.0020	0.0000

Figure 2.9: F_{st} values for the amplicon study. The coding sequences surrounded *pgFAR* are indicated as follows: dark blue, lipase3; red *pgFAR*; light green, vitellogenin; light blue, mitochondrial aldehyde dehydrogenase; orange, lipase; purple, mitochondrial sodium/hydrogen exchanger NHA2; pink, unnamed cds; dark green, Sfmbt-like polycomb protein.

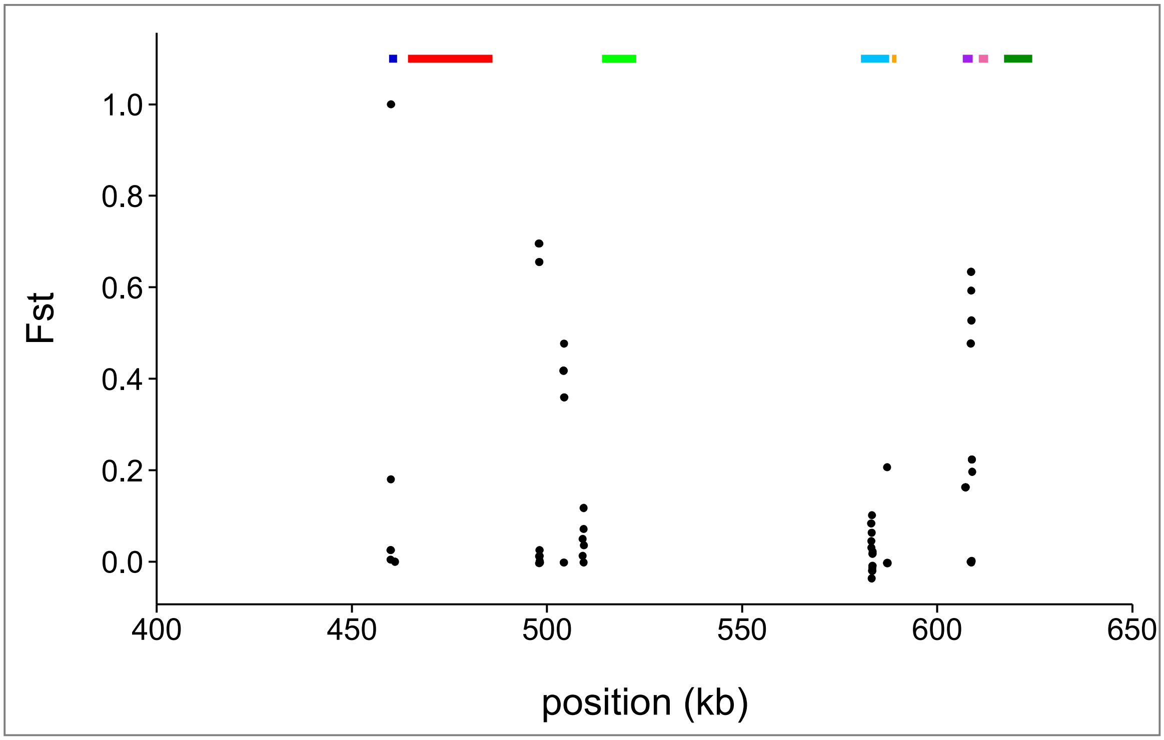
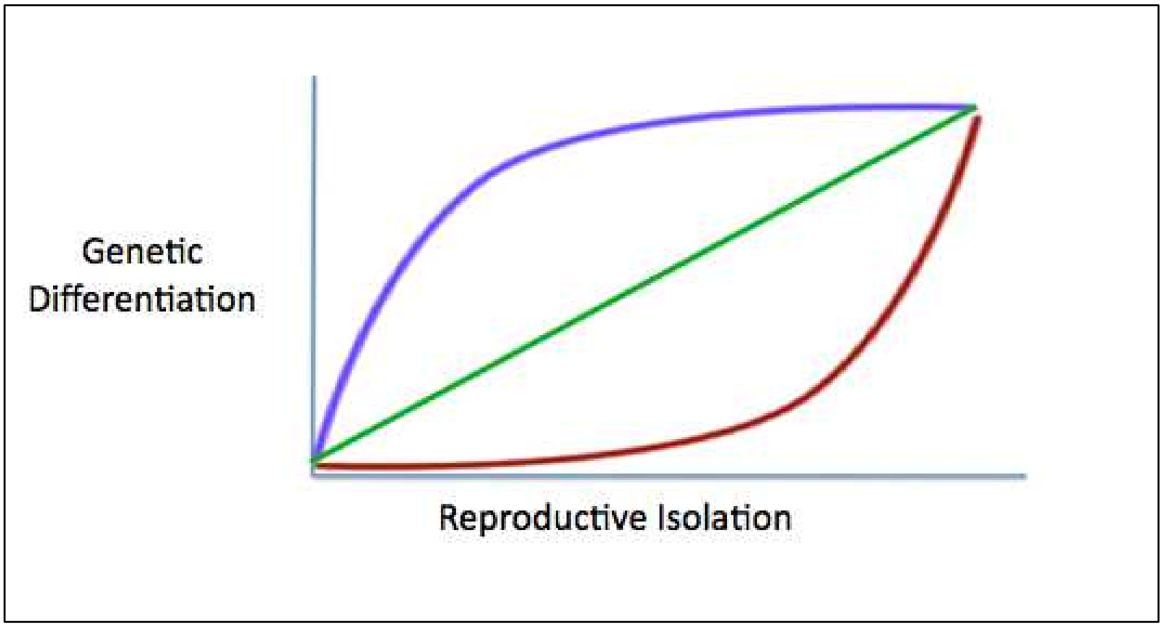


Figure 2.10: A model of the possible relationships between reproductive isolation and genetic differentiation. The green line represents a relationship in which genetic differentiation increases linearly with reproductive isolation. The purple line represents a relationship in which a small amount of reproductive isolation is sufficient to generate a large island of genetic differentiation, followed by limited impact of any additional reproductive isolation. The red line represents a relationship in which appreciable genetic differentiation is only generated under conditions of very strong reproductive isolation.



CHAPTER 3

Future Directions

In this project, we observed that a strong prezygotic reproductive isolating barrier is capable of generating a several kilobase long region of elevated genetic differentiation. In the future, we plan to extend this research by exploring the impact of various demographic factors on the size of the island of differentiation surrounding *pgFAR*. In particular, we plan to investigate the role of two different factors: 1) the presence of an additional reproductive isolating barrier elsewhere in the genome, 2) North American vs. European populations.

An additional reproductive isolating barrier:

In many locations, E and Z strain moths experience temporal isolation in addition to behavioral isolation. Both strains enter diapause as larvae in the fall and break diapause in the spring. However, the timing of the diapause break for E and Z strains can be substantially offset, resulting in sexually mature adults of the two strains which are present different times of the year (Glover *et al*., 1992). At intermediate latitudes, larvae that break diapause earlier in the year are able to squeeze a second generation in to the growing season (bivoltine), while larvae that emerge from diapause later can only fit in one generation (univoltine). In some locations, such as upstate New York, three different strains of corn borers are found: bivoltine E (BE), bivoltine Z (BZ), and univoltine Z (UZ) (Roelofs *et al.*, 1985).

A future project could explore the influence of temporal isolation as a second reproductive isolating barrier on the size of the island surrounding *pgFAR*. This would be accomplished by observing the region of differentiation around *pgFAR* for sympatric BE/BZ populations (for example, from Dover, Massachusetts) as well as sympatric BE/UZ populations (for example, from Geneva, New York). When both reproductive isolating barriers are present, gene flow will be reduced along the entirety of the genome compared to when only one barrier is present. This reduction in gene flow will reduce opportunities for recombination to introduce alleles from one population into the alternate genetic background. We predict that this overall reduction of gene flow will have two effects: 1) increase the height and length of island surrounding *pgFAR* and 2) increase the baseline F_{st} values in the "sea" (Figure 1). Together, these two effects may counter-intuitively lead to the appearance of a smaller island, if its size is judged in relation to the rising "sea" level (see Figure 1).

North American vs. European sympatric population pairs:

Another fruitful comparison would come from contrasting the size of the island of surrounding *pgFAR* in North American ECB populations with that of their European counterparts. There are several variables that might be expected to influence the size of the islands observed in population pairs from the two continents. First, while some European E and Z populations are thought to have been in sympatry for a long time, sympatry may be relatively recent in North American populations. This is because some of the Z individuals in North America originated in Hungary, where Z populations are found in allopatry. In general, we expect that for two populations which have returned to secondary contact after a period of allopatry, the longer the period of secondary contact, the smaller the island of divergence around reproductive isolating barriers will be (Figure 2). This is because gene flow and recombination between the two populations will have had more time to degrade the island surrounding the reproductive isolating barrier. Thus, we might expect to observe a narrower region of genetic differentiation in the sympatric European populations compared to the North American populations.

However, the comparison of North American and European ECB is further complicated by the possibility of a colonization effect. New selective pressures in recently colonized habitats might cause the region of differentiation around *pgFAR* to be bigger or smaller in North American populations, depending on the whether selection is divergent or not (Figure 3). This combination of forces generates three possible scenarios. If we were to observe similarly sized regions of differentiation in the population pairs from the two continents (Figure 3a), we would conclude that either 1) the colonizing individuals from Europe were primarily from sympatric populations or 2) that differences which accumulated in allopatry were not very stable and quickly eroded under gene flow. If we observed a larger region of differentiation in the North American population pair (Figure 3b), we might conclude that 1) the colonizing Z individuals did come from an allopatric source population and that 2) the differences that accumulated in allopatry do not degrade quickly. Another possible explanation of this pattern is that in the new environment, E and Z strains experience a differential selective pressure at a locus near *pgFAR*, which contributes to a larger island. Lastly, it is possible that we would observe a smaller region of differentiation in the North American population pair (Figure 3c). Such a pattern could be explained by a new selective pressure acting on a locus near *pgFAR* that homogenizes the two strains.

Both of these projects would expand our understanding of the factors that contribute to the accumulation of genetic differentiation around a reproductive isolating barrier under different ecological circumstances. Ultimately, these lines of study will allow us to better comprehend how genomes diverge during the speciation process.

FIGURES

Figure 3.1. A hypothesis about how the pattern of genetic differentiation surrounding *pgFAR* might differ in BE/BZ vs. BE/UZ population pairs. The red line represents genetic differentiation between sympatric BE and UZ populations, while the blue line represents genetic differentiation between sympatric BE and BZ populations. When both reproductive isolating barriers are present, the island of genetic differentiation surrounding *pgFAR* is predicted to be taller and wider and the "sea" level is also predicted to be higher.

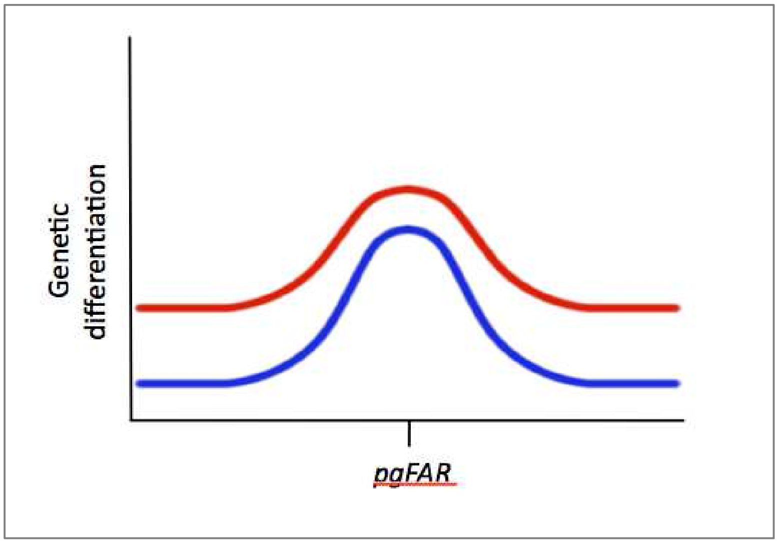


Figure 3.2. A hypothesis about how the region of differentiation around a reproductive isolating barrier would differ under three conditions: allopatry, recent sympatry and older sympatry. It is hypothesized that following secondary contact, gene flow and recombination would chip away at the region of differentiation, resulting in a narrower island.

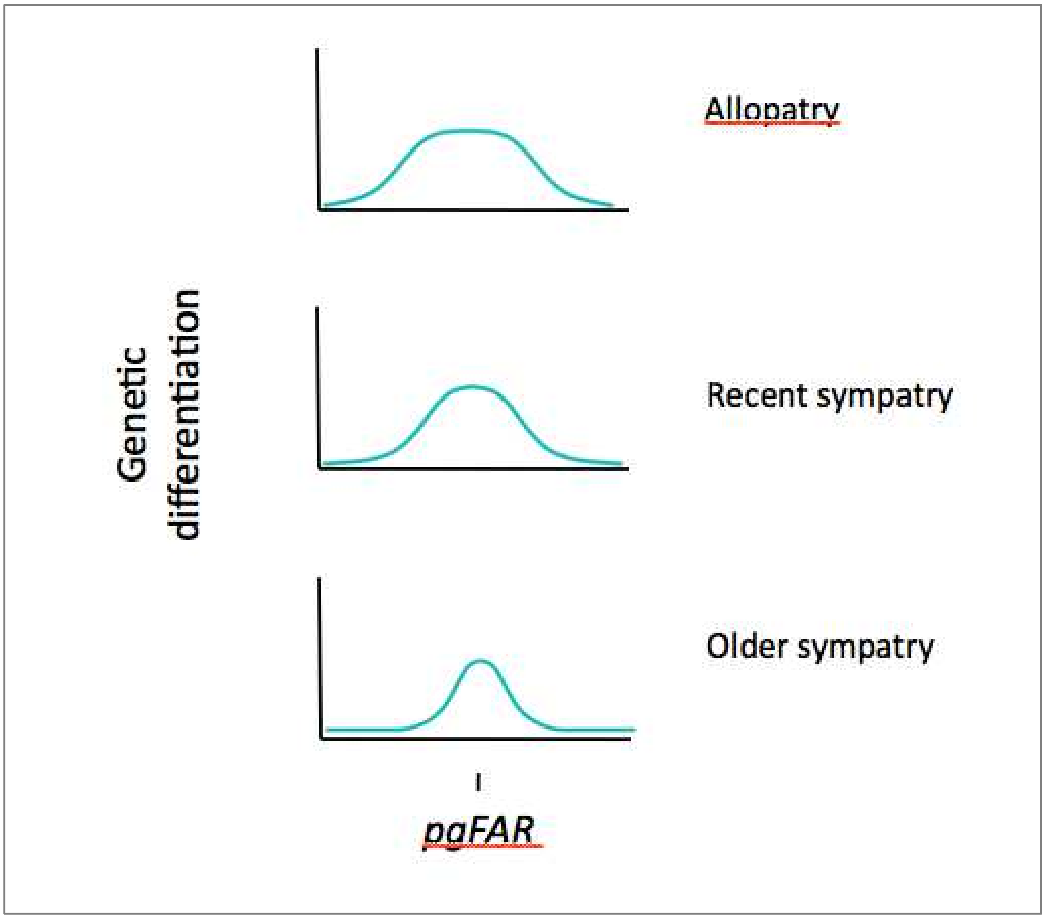
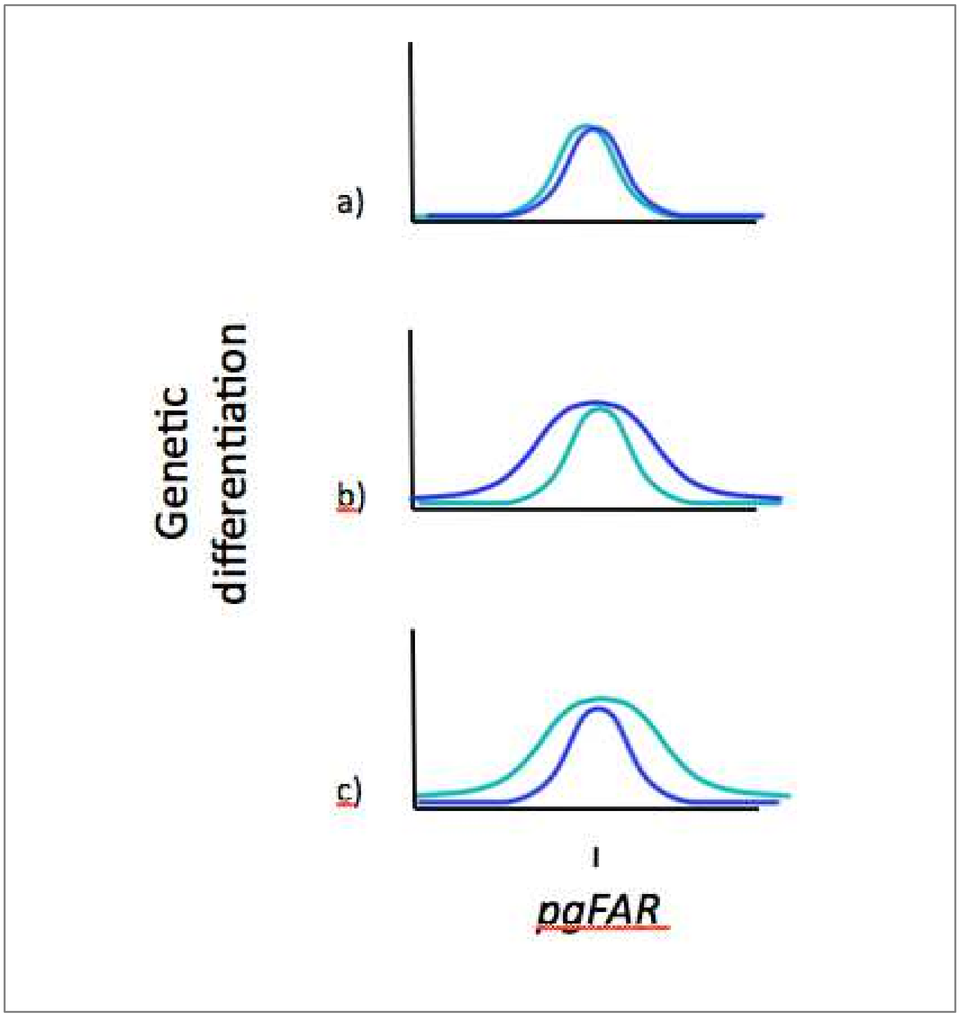


Figure 3.3. Three possible relationships between the size of the islands of genetic differentiation surrounding *pgFAR* in European and North American E and Z sympatric populations. Light blue line represents genetic differentiation between European populations and the dark blue line represents genetic differentiation between North American populations. Relationship a) might occur if the colonizing individuals from Europe were primarily from sympatric populations or that differences which accumulated in allopatry were not very stable. Relationship b) might occur if the colonizing Z individuals came from an allopatric source population and the differences that accumulated in allopatry do not degrade quickly. Relationship c) might occur if there is selection near the pgFAR island.



REFERENCES

Barker, J. S. F. (1962). Sexual isolation between Drosophila melanogaster and Drosophila simulans. *American Naturalist*, 105-115.

Barton, N. H. & Hewitt, G. M. (1981). Hybrid zones and speciation. In W. R. Atchley, D. S. Woodruff (Eds.), *Evolution and Speciation: Essays in Honour of M. J. D. White* (109-145). Cambridge: Cambridge University Press.

Baxter, S. W., Papa, R., Chamberlain, N., Humphray, S. J., Joron, M., Morrison, C., ... & Jiggins, C. D. (2008). Convergent evolution in the genetic basis of Müllerian mimicry in Heliconius butterflies. *Genetics*, *180*(3), 1567-1577.

Bradshaw, H. D., & Schemske, D. W. (2003). Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature*, *426*(6963), 176-178.

Brideau, N. J., Flores, H. A., Wang, J., Maheshwari, S., Wang, X. U., & Barbash, D. A. (2006). Two Dobzhansky-Muller genes interact to cause hybrid lethality in Drosophila. *Science*, 314(5803), 1292-1295.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bomblies, K., Lempe, J., Epple, P., Warthmann, N., Lanz, C., Dangl, J. L., & Weigel, D. (2007). Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biology*, *5*(9), e236.

Caffrey, D. J. & Worthley, L. H. (1927). A progress report on the investigations of the European corn borer. USDA Bull. No. 1476, Governmental Printing Office, Washington, DC.

Clarkson, C. S., Weetman, D., Essandoh, J., Yawson, A. E., Maslen, G., Manske, M., ... & Donnelly, M. J. (2014). Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nature Communications*, *5*, 4248.

Coates, B. S., Johnson, H., Kim, K. S., Hellmich, R. L., Abel, C. A., Mason, C., & Sappington, T. W. (2013). Frequency of hybridization between Ostrinia nubilalis E-and Z-pheromone races in regions of sympatry within the United States. *Ecology and evolution*, *3*(8), 2459-2470.

Coyne, J. A., & Orr, H. A. (2004). *Speciation* (Vol. 37). Sunderland, MA: Sinauer Associates.

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133-3157.

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., ... & Palumbi, S. R. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, *12*(6), 1058-1067.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491-498.

Dopman, E. B. (2011). Genetic hitchhiking associated with life history divergence and colonization of North America in the European corn borer moth. *Genetica*, *139*(5), 565-573.

Dopman, E. B., Bogdanowicz, S. M., & Harrison, R. G. (2004). Genetic mapping of sexual isolation between E and Z pheromone strains of the European corn borer (Ostrinia nubilalis). *Genetics*, *167*(1), 301-309.

Dopman, E. B., Pérez, L., Bogdanowicz, S. M., & Harrison, R. G. (2005). Consequences of reproductive barriers for genealogical discordance in the European corn borer. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(41), 14706-14711.

Dopman, E. B., Robbins, P. S., & Seaman, A. (2010). Components of reproductive isolation between North American pheromone strains of the European corn borer. *Evolution, 64*(4), 881-902.

Du, M., Yin, X., Zhang, S., Zhu, B., Song, Q., & An, S. (2012). Identification of lipases involved in PBAN stimulated pheromone production in Bombyx mori using the DGE and RNAi approaches. *PloS one*, *7*(2), e31045.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., ... & Wolf, J. B. (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, 491(7426), 756-760.

Ferguson, L., Lee, S. F., Chamberlain, N., Nadeau, N., Joron, M., Baxter, S., ... & Jiggins, C. (2010). Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. *Molecular Ecology*, *19*(s1), 240-254.

Franchini, P., Fruciano, C., Spreitzer, M. L., Jones, J. C., Elmer, K. R., Henning, F., & Meyer, A. (2014). Genomic architecture of ecologically divergent body shape in a pair of sympatric crater lake cichlid fishes. *Molecular Ecology*, 23(7), 1828-1845.

Fuster, D. G., Zhang, J., Shi, M., Bobulescu, I. A., Andersson, S., & Moe, O. W. (2008). Characterization of the sodium/hydrogen exchanger NHA2. *Journal of the American Society of Nephrology*, *19*(8), 1547-1556.

Glover, T. J., Tang, X. H., & Roelofs, W. L. (1987). Sex pheromone blend discrimination by male moths from E and Z strains of European corn borer. *Journal of Chemical Ecology*, *13*(1), 143-151.

Glover, T. J., Robbins, P. S., Eckenrode, C. J., & Roelofs, W. L. (1992). Genetic control of voltinism characteristics in European corn borer races assessed with a marker gene. *Archives of Insect Biochemistry and Physiology*, *20*(2), 107-117.

Gagnaire, P. A., Pavey, S. A., Normandeau, E., & Bernatchez, L. (2013). The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by rad sequencing. *Evolution*, 67(9), 2483-2497.

Goudet, J. (2005). HierFstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184-186.

Harrison, R. G. (1990). Hybrid zones: windows on evolutionary process. *Oxford Surveys in Evolutionary Biology*, *7*, 69-128.

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 Genetics*, *6*(2), e1000862.

Hopkins, R., & Rausher, M. D. (2011). Identification of two genes causing reinforcement in the Texas wildflower Phlox drummondii. *Nature*, *469*(7330), 411-414.

Hutchison, W. D., Burkness, E. C., Mitchell, P. D., Moon, R. D., Leslie, T. W., Fleischer, S. J., ... & Raun, E. S. (2010). Area wide suppression of European corn borer with Bt maize reaps savings to non-Bt maize growers. *Science*, *330*(6001), 222-225.

Jiggins, C. D., Mavarez, J., Beltrán, M., McMillan, W. O., Johnston, J. S., & Bermingham, E. (2005). A genetic linkage map of the mimetic butterfly Heliconius melpomene. *Genetics*, *171*(2), 557-570.

Jiggins, C. D., Naisbit, R. E., Coe, R. L., & Mallet, J. (2001). Reproductive isolation caused by colour pattern mimicry. *Nature*, *411*(6835), 302-305.

Kliman, R. M., Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A. J., ... & Hey, J. (2000). The population genetics of the origin and divergence of the Drosophila simulans complex species. *Genetics*, *156*(4), 1913-1931.

Kochansky, J., Cardé, R. T., Liebherr, J., & Roelofs, W. L. (1975). Sex pheromone of the European corn borer, Ostrinia nubilalis (Lepidoptera: Pyralidae), in New York. *Journal of Chemical Ecology*, *1*(2), 225-231.

Kofler, R., Orozco-ter Wengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., ... & Schlötterer, C. (2011a). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS one*, *6*(1), e15925.

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011b). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*(24), 3435-3436.

Kronforst, M. R., Young, L. G., Kapan, D. D., McNeely, C., O'Neill, R. J., & Gilbert, L. E. (2006). Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proceedings of the National Academy of Sciences*, *103*(17), 6575-6580.

Lachaise, D., David, J. R., Lemeunier, F., Tsacas, L., & Ashburner, M. (1986). The reproductive relationships of Drosophila sechellia with D. mauritiana, D. simulans, and D. melanogaster from the Afrotropical region. *Evolution*, 262-271.

Lassance, J. M., Groot, A. T., Liénard, M. A., Antony, B., Borgwardt, C., Andersson, F., ... & Löfstedt, C. (2010). Allelic variation in a fatty-acyl reductase gene causes divergence in moth sex pheromones. *Nature*, 466(7305), 486-489.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Linn, C. E., Young, M. S., Gendle, M., Glover, T. J., & Roelofs, W. L. (1997). Sex pheromone blend discrimination in two races and hybrids of the European corn borer moth, Ostrinia nubilalis. *Physiological Entomology*, *22*(3), 212-223.

Malausa, T., Leniaud, L., Martin, J. F., Audiot, P., Bourguet, D., Ponsard, S., ... & Dopman, E. (2007). Molecular differentiation at nuclear loci in French host races of the European corn borer (Ostrinia nubilalis). *Genetics*, *176*(4), 2343-2355.

Maynard Smith, J., & Haigh, J. (1974). The hitchhiking effect of a favourable gene. *Genetical research*, *23*(01), 23-35.

Mayr, E. (1942). *Systematics and the Origin of Species from the Viewpoint of a Zoologist*. Harvard University Press.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303.

Merrill, R. M., Gompert, Z., Dembeck, L. M., Kronforst, M. R., McMillan, W. O., & Jiggins, C. D. (2011). Mate preference across the speciation continuum in a clade of mimetic butterflies. *Evolution*, *65*(5), 1489-1500.

Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C., & Forejt, J. (2009). A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science*, *323*(5912), 373-375.

Nadeau, N. J., Whibley, A., Jones, R. T., Davey, J. W., Dasmahapatra, K. K., Baxter, S. W., ... & Jiggins, C. D. (2012). Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 343-353.

Noor, M. A., & Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, 103(6), 439-444.

Nosil, P. (2012). Ecological speciation. Oxford University Press.

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 332-342.

Nosil, P., & Schluter, D. (2011). The genes underlying the process of speciation. *Trends in Ecology & Evolution*, *26*(4), 160-167.

Nunes, M. D., Orozco-ter Wengel, P., Kreissl, M., & Schlötterer, C. (2010). Multiple hybridization events between Drosophila simulans and Drosophila mauritiana are supported by mtDNA introgression. *Molecular Ecology*, *19*(21), 4695-4707.

Orr, H. A., & Presgraves, D. C. (2000). Speciation by postzygotic isolation: forces, genes and molecules. *BioEssays*, 22(12), 1085-1094.

Palmer, D. F., Schenk, T. C., & Chiang, H. C. (1985). Dispersal and voltinism adaptation of the European corn borer in North America, 1917-1977. Minnesota Agricultural Experiment Station. Retrieved from the University of Minnesota Digital Conservancy, http://purl.umn.edu/139529.

Perez, D. E., & Wu, C. I. (1995). Further characterization of the Odysseus locus of hybrid sterility in Drosophila: one gene is not enough. *Genetics*, *140*(1), 201-206.

Phifer-Rixey, M., Bomhoff, M., & Nachman, M. W. (2014). Genome-wide patterns of differentiation among house mouse subspecies. *Genetics*, *198*(1), 283-297.

Presgraves, D. C., Balagopalan, L., Abmayr, S. M., & Orr, H. A. (2003). Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. *Nature*, 423(6941), 715-719.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Renaut, S., Maillet, N., Normandeau, E., Sauvage, C., Derome, N., Rogers, S. M., & Bernatchez, L. (2012). Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 354-363.

Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., ... & Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, *4*, 1827.

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, *16*(7), 351-358.

Roelofs, W. L., Du, J. W., Tang, X. H., Robbins, P. S., & Eckenrode, C. J. (1985). Three European corn borer populations in New York based on sex pheromones and voltinism. *Journal of Chemical Ecology*, *11*(7), 829-836.

Ruegg, K., Anderson, E. C., Boone, J., Pouls, J., & Smith, T. B. (2014). A role for migration-linked genes and genomic islands in divergence of a songbird. *Molecular Ecology*, *23*(19), 4757-4769.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585-595.

Tamura, K., Subramanian, S., & Kumar, S. (2004). Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Molecular Biology and Evolution*, *21*(1), 36-44.

Tang, S., & Presgraves, D. C. (2009). Evolution of the Drosophila nuclear pore complex results in multiple hybrid incompatibilities. *Science*, 323(5915), 779-782.

Ting, C. T., Tsaur, S. C., Wu, M. L., & Wu, C. I. (1998). A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science,* 282(5393), 1501-1504.

Ting, C. T., Tsaur, S. C., & Wu, C. I. (2000). The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. *Proceedings of the National Academy of Sciences*, *97*(10), 5313-5316.

True, J. R., Mercer, J. M., & Laurie, C. C. (1996). Differences in crossover frequency and distribution among three sibling species of Drosophila. *Genetics*, *142*(2), 507-523.

Turner, T. L., & Hahn, M. W. (2010). Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, *19*(5), 848-850.

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in Anopheles gambiae. *PLoS Biology*, *3*(9), 1572-1578.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., ... & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 11:11.10:11.10.1–11.10.33.

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 451-460.

Via, S., & West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, *17*(19), 4334-4345.

Wadsworth, C. B., Li, X., & Dopman, E. B. (2015). A recombination suppressor contributes to ecological speciation in OSTRINIA moths. *Heredity*, *114*(6), 593-600.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256-276.

Wessinger, C. A., Hileman, L. C., & Rausher, M. D. (2014). Identification of major quantitative trait loci underlying floral pollination syndrome divergence in Penstemon. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1648): 20130349.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer Science & Business Media.

Wilfert, L., Gadau, J., & Schmid-Hempel, P. (2007). Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity*, *98*(4), 189-197.

Wu, C. I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851-865.

Wu, C. I., & Ting, C. T. (2004). Genes and speciation. *Nature Reviews Genetics*, 5(2), 114-122.

Wu, S., Trievel, R. C., & Rice, J. C. (2007). Human SFMBT is a transcriptional repressor protein that selectively binds the N-terminal tail of histone H3. *FEBS letters*, *581*(17), 3289-3296.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, *13*(1), 134-144.

You, F. M., Huo, N., Gu, Y. Q., Luo, M. C., Ma, Y., Hane, D., ... & Anderson, O. D. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*, *9*(1), 253-265.

Zhang, L., Sun, T., Woldesellassie, F., Xiao, H., & Tao, Y. (2015). Sex Ratio Meiotic Drive as a Plausible Evolutionary Mechanism for Hybrid Male Sterility. *PLoS Genetics*, *11*(3) e1005073-e1005073.

Questions from the public presentation:

1. Are there other reproductive isolating barriers that you could query using your whole genome sequencing? (you mentioned there were 7 in your intro).

2. You mentioned that pi and theta remain lower in the pgFAR locus in the E and Z populations because there is less ability to maintain diversity in this locus in the two populations. What would be the molecular basis for this?

3. What do you predict the Fst plot will look like to the right of position 650 kb? Do you think you've found the entire island?