Population differentiation and structural variation in the genome of *Manduca sexta* across the United States

Andrew J. Mongue[1,2,*], Akito Y. Kawahara[3]

[1] University of Edinburgh, Institute of Evolutionary Biology, Charlotte Auerbach Road, Edinburgh, UK, EH93FL

[2] University of Kansas, Department of Ecology and Evolutionary Biology, 1200 Sunnyside Ave, Lawrence, KS, USA, 66045

[3]McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL 32611

[*]correspondence contact: amongue@ed.ac.uk

Abstract

Many species that are extensively studied in the laboratory are less well characterized in their natural habitat, and laboratory strains represent only a small fraction of the variation in a species' genome. Here we investigate genomic variation in three natural populations of an agricultural pest and a model insect for many scientific disciplines, the tobacco hornworm (*Manduca sexta*). We show that hornworms from Arizona, Kansas, and North Carolina are genetically distinct, with Arizona being particularly differentiated from the other two populations. Specifically, two segregating inversions and a potential pseudogene are found only in the Arizona population. One inversion on the Z chromosome may enhance adaptive evolution of the sex chromosome, while the significance of the other, autosomal inversion remains unclear. The pseudogene may be involved in the exploitation of a novel hostplant in Arizona, but functional genetic assays will be required to confirm this hypothesis. Nevertheless, our results reveal undiscovered natural variation and provide useful genomic data for a model insect species.

Introduction

One of the main objectives of evolutionary genetics is to unite the proximate molecular function of genes with the long-term evolutionary forces that govern their change (Nei, 1987). The two halves of this pursuit necessitate different approaches, with well-controlled laboratory studies best suited to elucidating the biochemistry of gene function and surveys of natural variation required to understand the ecological and demographic context in which these genes exist. The most fruitful species to study in the lab tend to be those with small body sizes and short generation times. In the field, common species with large population sizes are easier to sample. Consequently, many of the most popular models for evolutionary genetics come from insects, which fulfill all of the above criteria.

The tobacco hornworm moth (*Manduca sexta*) is a well-established laboratory model (Reinecke, Buckner, & Grugel, 1980), and likely the second most studied moth after the domestic silkmoth, *Bombyx mori* (The International Silkworm Genome Consortium, 2008). The tobacco hornworm is used to study developmental biology from eyes to gonads (Banister & White, 1987; Höglund & Struwe, 1970; Nardi, 1993), immunology (Kanost, Jiang, & Yu, 2004), reproductive biology (Mongue & Walters, 2017;

Whittington, Zhao, Borziak, Walters, & Dorus, 2015), along with a range of other topics. This attention has resulted in a wealth of genetic information for this species, with many physiological functions characterized by the genetic pathways on which they depend and supported by robust sequencing of RNA transcripts (aggregated in Cao & Jiang, 2017). In other words, there has been great progress in understanding the molecular genetics of this species.

Yet for all of this research in vitro, the genetics of *M. sexta* in its natural habit remains largely unexplored. Its phylogenetic relationship to other sphinx moths was only recently genetically evaluated (Kawahara et al., 2013) and only last year the first whole genome sequences from wild individuals were generated (Mongue, Hansen, Gu, Sorenson, & Walters, 2019). Aside from these studies, very basic questions of natural populations have yet to be addressed. For instance, *M. sexta*, like other moths in the family Sphingidae are particularly strong and mobile fliers (Stevenson, Corbo, Baca, & Le, 1995) and at one time it was claimed that this species is migratory (Ferguson, 1991). However, both this initial claim and its later refutation are based solely on limited observations of seasonal occurrence (McNeil, 1995) and the ability of pupae to survive subfreezing temperatures (Shapiro, 2006). Surprisingly, there have been no direct studies of population structuring of *M. sexta* to date.

This lack of research into the population genetics of this moth is puzzling on two fronts. First, in terms of basic research, much of the ecological knowledge about this moth, such as acclimation and adaptation to different environments and host plant usage (Contreras et al., 2013; Mechaber & Hildebrand, 2000), could benefit from understanding natural genetic variation across its range. Second, as the name tobacco hornworm suggests, *M. sexta* is a pest of economically important crops, including tobacco (*Nicotiana* spp.), tomatoes, peppers, and generally other plants in the nightshade family, Solanaceae (Wagner, 2005). Thus, understanding how *M. sexta* populations are structured will inform pest management strategies. In service of both of these goals, we present whole-genome resequencing of this model insect to investigate population structure across an east-west transect of the continental United States. With these resequencing samples, we establish patterns of regional genetic variation and differentiated parts of the genome for further studies of *M. sexta.*

Methods

Sample collection and sequencing

Samples from three geographic regions in the US were examined in this project: North Carolina, Kansas, and Arizona. Collecting methods varied somewhat between these locales. Samples from North Carolina were already obtained and sequenced as part of Mongue *et al.* (2019). In brief, adults were collected over the course of a week from a mercury vapor light trap in a single locale. Kansas samples were obtained from a mixture of adult moths caught at light traps and larvae taken from tomato fields in different towns in eastern Kansas. For larval sampling, only one individual was sequenced per field to minimize the risk of sequencing siblings. Arizona samples were obtained from the University of Florida's collections, where adult moths had been preserved in ethanol. At time of collection, they were obtained from a mercury vapor trap at one location over several nights. A summary of sequenced individuals and their methods of collection can be found in Table S1.

For each sample, tissue was extracted from adult thoracic flight muscle to avoid eggs and sperm, and larval head capsules to avoid hostplant tissue in the gut. DNA was extracted using an Omniprep kit for

genomic DNA (G-Biosciences, St. Louis, MO, USA), following manufacturer's protocol. From these extractions, 150 basepair, paired-end Illumina reads were sequenced to an average depth of 14-20x coverage across the genome (also summarized in Table S1).

Short-read data were aligned to the *M. sexta* reference genome (Kanost et al., 2016) using the very sensitive local aligner in Bowtie2 version 2.2.9 (Langmead & Salzberg, 2012). Alignments were sorted and optical duplicates were removed with Picardtools v. 2.8 (Wysoker, Tibbetts, & Fennell, 2013). Finally, alignments around insertions and deletions were adjusted with GATK v. 3.7 (McKenna et al., 2010). These curated alignments were the starting point for all downstream analyses.

Population structure

For population structuring, alignments were filtered to separate the Z chromosome, as unequal sampling of sexes between populations could create spurious patterns. Chromosomal linkage information came from previous synteny analyses comparing *M. sexta* to *B. mori* to identify and separate sex-linked scaffolds for these analyses (Mongue & Walters, 2017). The autosomal genome was used to infer population structure while accounting for uncertainty in genotyping using the software NGSAdmix v. 3.2 (Skotte, Korneliussen, & Albrechtsen, 2013) as well as a principle-component based approach (Meisner & Albrechtsen, 2018). For the admixture analysis, the best-supported number of populations (K) was determined by assessment of ten bootstrapped replicates each for possible K values from 2 to 6, evaluated using CLUMPAK's web-hosted application (Kopelman, Mayzel, Jakobsson, Rosenberg, & Mayrose, 2015).

Characterization of genomic differences

After establishing basic population relationships, population differences were examined in detail. A baseline expectation of differentiation was set by estimating the genome-wide pairwise F_{ST} (excluding the Z chromosome) analyzing the curated alignments with the population genetic software ANGSD v. 0.916 (Korneliussen, Albrechtsen, & Nielsen, 2014). This global estimate was performed separately for putatively neutral (four-fold degenerate) and selected (zero-fold degenerate) sites to assess whether differentiation skewed towards neutral or adaptive differentiation. Identification of zero-fold and four-fold degenerate sites was completed in previous analyses with custom R scripts (Mongue et al., 2019; R Core Team, 2017).

With a baseline established, candidates for further investigation were identified via a preliminary scan for F_{ST} outlier regions in windows of 10 kilobases with a 5 Kb step size to identify regions (i.e. scaffolds) for further study. All pairwise comparisons were performed: North Carolina (NC) – Arizona (AZ), Kansas (KS) – Arizona, and NC – KS. For these analyses, published chromosome assignments of scaffolds were employed (Kanost et al., 2016; Mongue, Hansen, & Walters, 2020; Mongue & Walters, 2017). Importantly, these assignments, which used a mixture of synteny (i.e conserved ordering of orthologous genes) and sex-differences in coverage (to identify the Z chromosome), did not contain information on ordering of scaffolds *within* a chromosome.

An attempt was made to reconstruct the order of scaffolds on the relevant chromosome. To do so, the *M. sexta* reference was aligned to the chromosome level assembly of *Bombyx mori* (Kawamoto et al., 2019), because the two species have been shown to have very conserved genome structure (Yasukochi et al., 2009). Syntenic alignments were carried out using Satsuma (Grabherr et al., 2010). Another line of evidence came from a recent pre-publication of a near-chromosomal reference assembly of *M. sexta*

(Gershman et al., 2020). This assembly has not undergone peer review to vet its quality and it lacks chromosomal assignment of scaffolds; thus, it was not chosen for full the initial genome-wide analyses. Nonetheless, the more contiguous scaffolds could be useful in confirming the number of putative inversions observed as well as anchoring unplaced scaffolds to chromosomes. To do so, another syntenic alignment with Satsuma v. 3.1 (Grabherr et al., 2010) was performed, this time between the two *M. sexta* assemblies. For relevant scaffolds, alignment to this new reference and calculation of F_{ST} between North Carolina and Arizona followed the methods described above.

Finally, to characterize which population(s) carried the ancestral or derived orientation in the putatively inverted region(s), genetic diversity was examined in these regions in two ways. First ANGSD (Korneliussen et al., 2014) was employed to estimate pairwise genetic differences (π) along chromosomes of interest for each population. π in one population was subtracted from π in the same region of another population for each of the three pairwise comparisons. These difference values were plotted along the chromosome with the neutral expectation that two populations should not substantially differ in genetic diversity outside of inverted regions. Second, single nucleotide polymorphism (SNP) frequencies were examined in each population, under the logic that loci within an inversion cannot freely recombine and thus should share frequencies across longer distances. To this end, Variant Call Format (vcf) files were generated with GATK's best practices pipeline (McKenna et al., 2010), including hard quality filters (Quality by Depth > 2.0 & Fisher Strand-bias < 60 & Mapping Quality > 40). From these vcfs, alternate allele frequencies were extracted and examined. Overall frequencies are plotted in the results, but the VCFtools utility "extract-FORMAT-info" (Danecek et al., 2011) was used to examine sample genotypes directly and determine which individuals carried non-reference alleles as well.

Genetic characterization of differentiated regions

For each genomic region-of-interest, annotated genes in the region were examined. *Manduca sexta* has a wealth of available RNA sequence data from both sexes and different life-stages, aggregated as a set of RNAseq collected by Cao and Jiang (2017). From this dataset, the specificity of gene expression was calculated for each gene in the genome for life-stage, tissue, and sex-specificity (following methods in Mongue, Hansen, & Walters, 2020) to inform potential function of genes-of-interest. More precisely, the specificity metric SPM was used to assess the proportion of a given gene's expression found in a specific class. This metric ranges from zero to one, inclusive, and denotes the proportion of a given gene's expression in a focal category (Kryuchkova-Mostacci & Robinson-Rechavi, 2017). SPM was calculated for each of the three dimensions of specificity. Stage-specificity was defined as gene expression unique to larva, pupa, or adult *M. sexta* based on a set of tissues collected in all three stages (head, midgut, and fat bodies). Tissue-specificity within a stage was based on tissue types available for that stage (see Table 1). And, naturally, sex-specificity was based on amounts of expression unique to males or females. For instance, a gene with an adult SPM of 0.90 shows 90% of its expression in adults as opposed to pupae or larvae. The same gene has separate SPM values that denote how specific its expression is to males or females and which tissues of those individuals. With this method has the benefit of generating a sense of gene-function independent of functional annotation based on sequence homology with genes of known function.

All genes across the genome were annotated with SPM values; however, for large outlier regions spanning numerous genes, it was not feasible to examine individual genes in depth. Instead, using the above annotations, these regions were examined for biased composition compared to non-differentiated regions. In other words, enrichment or depletion of functional classes within inversions was tested. For small regions containing one or a few genes, each could be manually examined. In these cases, single nucleotide polymorphisms in and around these genes were examined for evidence of coding-sequence variation between populations. SNPs were annotated with the tool SNPeff v. 4.2, which uses gene models to predict the coding sequence changes introduced by polymorphisms (e.g. gain of a premature stop codon, Cingolani et al., 2012).

Results

Structure

We found three populations via the method of Evanno et al. (2005): namely, the most likely K is the one the biggest change in likelihood from the previous K value. Per this metric, each of the three states' moths form distinct populations, although there appears to be some admixture between Arizona and Kansas, including an Arizona individual confidently grouping with the Kansas samples (Figure 1a).

Taking a complimentary approach, we assessed principle components of SNP variants across our samples and observed the same clustering pattern as found in the structure analysis. Each state's samples do not cluster with other states, but Arizona samples showing a more dispersed pattern than the others, including some individuals more similar to Kansas individuals than other Arizona individuals (Figure 1b).

Differentiation across the genome

A simple genome-wide average of F_{ST} set a baseline level of differentiation of 0.04 to 0.09, depending on the pairwise comparison between states (Table 2). A sliding window scan across the genome revealed numerous scaffolds with markedly higher F_{ST} (> 0.5, Figure 2). Using previously established synteny with *Bombyx mori* (Mongue & Walters, 2017), we placed most of these outliers onto chromosomes. Note however, that this level of macro-synteny only revealed chromosomal linkage groups, *not* ordering of scaffolds within chromosomes. In other words, multiple F_{ST} peaks on a given chromosome could represent independent differentiated regions or extended tracts of linked differentiation. Additionally, although synteny is generally well-conserved between *Manduca* and *Bombyx*, some regions have diverged enough between species that they could not be placed on scaffolds via synteny-approaches. For Z-linked regions, we could recover linkage of non-syntenic scaffolds based on male-to-female coverage differences (Mongue et al., 2020) but for autosomal scaffolds, these sequences remain unplaced for the time being.

From this analysis, several regions-of-interest became apparent. Most notably, a large proportion of chromosome 12 shows high levels of differentiation between Arizona and the other two populations. The Z chromosome showed the strongest peak of elevated differentiation, and single peaks were identified on chromosomes 2, 6, and 13, as well as two on chromosome 19. In the following sections we describe each of these in detail.

A segregating inversion on chromosome 12

We observed a large cluster of similar F_{ST} peaks across entire scaffolds, all belonging to chromosome 12. These peaks correspond to 13 scaffolds, listed in Table S2, summing to 7.6 Mb of sequence, approximately 56% of the total (assembled) length of chromosome. The peaks contain 321 annotated genes. The whole-scaffold nature of differentiation and between population patterns suggested two things: firstly, that the multiple signals could be the result of one (or a few) large-scale genomic variants (e.g. inversion(s)), and secondly, that whatever the variant(s), KS and NC were more similar to each other, with AZ being different.

To investigate further, we generated syntenic alignments with *B. mori* to order these scaffolds according to the arrangement of chromosome 12 in *B. mori*. This ordering resolved the numerous peaks into two regions of extended differentiation, though some smaller scaffolds could not be placed (Figure 3, top). These results suggested that the multiple peaks were not independent but raised the question of whether they truly were two separate blocks or one larger inversion.

To answer this question, we used a newly completed, near-chromosomal assembly of *M. sexta*. We aligned our short-read data from North Carolina and Arizona to this new reference and completed another F_{ST} scan, with a particular interest in chromosome 12. Based on this more contiguous assembly, we see that the differentiation belongs to one long tract of 8.1 Mb (Figure 4, Top). The increase in inversion size comes from two previously unplaced outlier scaffolds which we identified as belonging to chromosome 12 based on this alignment (shown corrected in Figure 2). Thus, we concluded that a single large inversion is responsible for the differentiation observed on this chromosome. The apparent splitting of this feature when using synteny with *Bombyx* is likely the result of fixed rearrangements that have occurred between these two species.

Finally, we examined which population(s) carried this inversion by examining frequencies of SNPs in each of the three populations (Figure 4, Bottom). We found that SNPs in the Arizona population have highly consistent frequencies across the elevated regions, most commonly at 0.6875. Conversely SNP frequencies were more variable across this region in North Carolina and Kansas, with rare variants the most common, as expected. This result is congruent with the d difference in π observed in Figure 3; pairwise differences (π) are maximized at intermediate allele frequencies and lower when an allele is very rare or very common. The consistent allele frequencies across the length of the putative inversion suggests that recombination is suppressed only in the westernmost population. In other words, the Arizona population carries the segregating inversion.

The inversion frequency of 0.6875 corresponds to 11 of 16 sampled chromosomes with the inversion. Per investigation of sample genotypes, this variation is distributed as 4 individuals homozygous for the inversion (A36, A70, A78, A82), three heterozygous for it (A71, A76, A84), and one homozygous for the ancestral orientation (A85). Note that the heterozygous individuals correspond to those inferred to be admixed with Kansas and the one homozygous non-inverted individual was grouped with Kansas completely in structure analyses. This pattern suggests that much of the signal for differentiation in Arizona comes from the rearrangement of chromosome 12.

The Z holds another segregating inversion

The Z chromosome's outliers are attributable to one (scaffold00022). Finer-scale plotting of differentiation reveals elevated F_{ST} extends across a ~1.4 Mb region of scaffold00022. The scale of

242 elevated differentiation suggests another inversion, and alignment to the newer *M. sexta* assembly
243 places the peak from scaffold00022 in the middle of the Z chromosome (HiC_scaffold_19) as a roughly
244 1.4 Mb region, suggesting we are capturing essentially all of the inversion on one scaffold, even in the
245 more fragmented assembly.

246 Examining allele frequencies, scaffold00022 shows an extended tract of SNPs at frequencies 0.875 and
247 1.0, which correspond to 7 of 8 alleles and all 8 of 8 respectively (Figure 5, Bottom). So, as with the
248 inversion on chromosome 12, this variant seems to be unique to the Arizona population.

249 In terms of gene content, scaffold00022 contains 35 genes and shows an significant excess of unbiased
250 genes (i.e. those expressed in both sexes, $X^2_2 = 11.36$, $p = 0.0034$), compared to the rest of the Z, which
251 is significantly masculinized, as previously reported (Mongue et al., 2020).

252

253 A potential pseudogene on Chromosome 13

254 The observed F_{ST} peak for this chromosome comes from one scaffold (scaffold00381) beyond 200 Kb
255 (Figure 6a). This region contains a single gene, Msex2.10493, and it displays a striking pattern of
256 differentiation. Within population variation is low in Kansas and North Carolina, with only 1 and 3
257 coding-sequence polymorphisms observed, respectively; in contrast, 61 SNPs were called from the
258 Arizona samples (summarized in Figure 6b). Moreover, many of these SNPs are at high frequency (> 0.5,
259 Figure 6c), including the predicted loss of a stop codon.

260 The massive difference in called-variants theoretically could arise from population-specific differences in
261 mapping rates in this region, either through poor mapping in populations with low SNP counts or (far
262 less likely) a pile-up of reads in Arizona owing to a collapsed segregating paralog. To rule out these
263 artefactual explanations, we examined per-base coverage around this gene and found no differences
264 between populations (Figure 6d; note that slightly deeper coverage in NC is consistent with genome-
265 wide differences in sequencing effort between the populations).

266 Thus, Arizona *M. sexta* appear to truly differ from the other populations with regard to the molecular
267 evolution of this gene, which begs the question of the gene's functional role. Msex2.10493 is
268 overwhelmingly (98.2%) expressed in larvae compared to 1.8% adult expression and vanishingly small
269 pupal expression. Within larvae, this gene shows 97% of its expression in the antennae and another
270 2.8% in the head.

271 Chromosomes 2, 6, and 19 show differentiation but no clearly differentiated genes

272 Chromosome 2 has a comparatively simple spike in F_{ST} at the end of scaffold00039, spanning roughly
273 50Kb of sequence, containing two annotated genes (Msex2.02572 and Msex2.02573); however, highly
274 differentiated SNPs are localized to areas outside the coding region of either. These variants may be
275 truly neutral or impact the regulation of the nearby genes, but without a better understanding of
276 regulatory elements (e.g. promoter sequences) in this species, it is impossible to say.

277 Chromosome 6 shows a single peak of divergence on scaffold00612, though this region is only weakly
278 differentiated, barely meeting our cut-off threshold for exploration ($F_{ST} > 0.5$ between any two
279 populations). This region contains three full annotated genes (Msex2.12148, Msex2.12149, and
280 Msex2.12150), but again no SNPs within coding regions.

Two outlier peaks localize to chromosome 19; both represent differentiation between Arizona and the other two states. One is localized to scaffold00118, to approximately the last 400Kb of an ~850Kb scaffold. This region contains 11 annotated genes (Msex2.05775-Msex2.05786), but none of these shows skews in coding region SNPs between populations as extreme as seen on chromosome 3. However, two genes do show more modest deviation between states. Msex2.05777 is variable in both North Carolina (Pn = 6, Ps = 10) and Kansas (Pn = 7, Ps = 12) but monomorphic (Pn = 0, Ps = 0) in Arizona. Conversely, Msex2.05784 is non-variable in the Southeast and Midwest but contains 3 non-synonymous variants and 1 synonymous variant in Arizona.

The second peak, on scaffold00419, seems to span the length of this 200Kb scaffold, but this may reflect the limits of resolution of our F_{ST} window-size. In any case, the entire scaffold contains 11 annotated genes, many of which have no expression data and no called SNPs in any population. As such, we do not place much emphasis on the differentiation around these dubiously annotated regions.

Discussion

We conducted whole genome resequencing to investigate the natural variation of *Manduca sexta* populations across the United States. In brief, we found that individuals from three populations across an East-West transect are genetically differentiable. This result may not seem surprising, but studies of other Lepidoptera have at times found evidence for continent-wide panmixis (specifically for monarch butterflies, Lyons et al., 2012; Zhan et al., 2014). And while monarchs may be the most extreme case of lepidopteran long-distance travel, sphinx moths, including *M. sexta* are renowned for their flight and dispersal ability, both in terms of small-scale speed and dexterity (Stevenson et al., 1995) and long-distance dispersal (Haber & Frankie, 1989; Janzen, 1984) Indeed, *M. sexta* has been posited as migratory in the past (Ferguson, 1991). Our results, however, suggest that gene flow is less common than expected of a species with routine migration. In fact, the genome-wide mean F_{ST} for the least differentiated pair (0.04, Kansas-Arizona) is on par with the differentiation between migratory North American monarchs and the Hawaiian population that is thought to have diverged roughly 200 years ago (Lyons et al., 2012).

More striking than this level of differentiation is the nature of the differences that contribute to it. Namely, large (>1 Mb) inversions, are key contributors to the variation between populations. Sustained tracts of differentiation are apparent on two chromosomes, the Z chromosome and chromosome 12. In both cases, the Arizona population is distinct from Kansas and North Carolina. Below we discuss in detail the genomic and genetic nature of these differences, with the caveat that functional genetics will be required to truly confirm the organismal and fitness consequences of this variation.

The largest signal for population differentiation: chromosome 12's segregating inversion

Scaffolds belonging to chromosome 12 were the most common outliers in our F_{ST} scan. By examining a combination of synteny mapping, a more contiguous *M. sexta* assembly, and allele frequencies in each population, we determined that these signals all belonged to a single massive inversion, spanning 8 megabases (roughly half the length of the chromosome) and containing 321 genes and found at a frequency of almost 70% in the Arizona population. Individuals that are heterozygous for the inversion or lacking it entirely cluster more closely with Kansas population than other Arizona samples which are homozygous for it. At present, it is unclear if this variation is representative of long-term local diversity

320 or (the offspring of) recent migrants to the population. The latter case would imply that the local
321 frequency of the inversion is even higher than we observe here.

322 In general, and especially in large populations, inversion frequencies are thought to imply something
323 about their selective effects, with only beneficial variants reaching high frequency (Kirkpatrick & Barton,
324 2006; Lande, 1984). In other species in which naturally occurring inversions have been studied, this
325 expectation has been borne out. Low frequency inversions (e.g. 8%) are associated with fitness tradeoffs
326 (Lee, Fishman, Kelly, & Willis, 2016) and more common inversions (~50%) have little to no fitness costs
327 (Knief et al., 2016). Thus, the common chromosome 12 variant in Arizona *M. sexta* seems likely to be at
328 least selectively neutral, if not beneficial. However, the large number of encompassed genes present a
329 challenge for identifying selectively relevant loci from purely computational methods.

330 A Z inversion in AZ: a role for sex-bias driving rearrangement

331 The single highest peak of differentiation in our genome-wide scan comes from the Z chromosome.
332 Scaffold00022, with roughly 1.4 Mb of extended differentiation and shared allele frequencies in Arizona
333 is fits the pattern of a segregating inversion between populations. Moreover, it is nearly fixed in the
334 Southwest population, being homozygous (or present in hemizygous females) in all but one individual.
335 As discussed above, an inversion is unlikely to reach such high frequency without some selective
336 advantage.

337 Here, we have a better candidate for what could drive this inversion to near-fixation. We found that the
338 inversion contains a significant excess of genes expressed in both sexes (i.e. unbiased genes), compared
339 to the rest of the Z chromosome, which is male-biased (Mongue & Walters, 2017). As has been recently
340 shown, unbiased genes are an important source of positive selection on the Z in *M. sexta* (Mongue et al.,
341 2020). These genes are expressed in a haploid state in females (which have only one Z chromosome);
342 thus, even recessive mutations are exposed to selection (Charlesworth, Coyne, & Barton, 1987; Rice,
343 1984). The capture of positively selected variants of such genes in a non-recombining region would
344 ensure their co-transmission and seems a reasonable explanation for this inversion's prevalence.

345 Other differentiated regions of the genome and a potential pseudogene in Arizona

346 The inversions discussed above are the most apparent sources of regional differentiation, but
347 chromosomes 2, 6, 13, and 19 also showed elevations in differentiation at smaller scales. For the most
348 part, differentiated regions fell outside of coding regions or near dubiously annotated genes (i.e. those
349 not supported by expression data) and for brevity are not discussed in detail here. One peak of
350 differentiation, on chromosome 13, deserves special mention, however.

351 Differentiation on chromosome 13 comes from a strong but very localized peak that can be attributed to
352 a single gene, Msex2.10493. It carries almost no variation in North Carolina or Kansas but has a wealth
353 of high frequency variants (61 SNPs at frequencies > 0.6 in our sample) in the Arizona population, one of
354 which is a predicted stop-loss in the coding region. A large-effect SNP like this could easily break the
355 protein function via readthrough translation of otherwise non-coding DNA. While inconclusive from
356 molecular data alone, a hypothetical loss of function would easily explain the preponderance of SNPs
357 found in Arizona; pseudogenization instantly relaxes selective pressure such that variants can accrue
358 neutrally. And indeed, the high frequency of this stop-loss suggests that it cannot have an appreciable
359 fitness cost, or it would not have persisted in the population. Naturally, this raises the question of what
360 function this gene has that it could be lost without strong selective consequences.

Based on our expression characterization, the Msex2.10493 gene is almost exclusively expressed in larval antennae, which are known to play a role in olfaction and food choice in *M. sexta* (Boer & Hanson, 1987). Olfaction genes are among the most commonly observed to pseudogenize in other species, ostensibly because the fitness consequences of their loss are rarely lethal (Nei, Niimura, & Nozawa, 2008; Prieto-Godino et al., 2016). Intriguingly, while most populations of *M. sexta* feed almost exclusively on plants in the family Solanaceae, Arizona hornworms have been recorded eating the unrelated devil's claw (Martyniaceae: *Proboscidea parviflora*, Mechaber & Hildebrand, 2000). It is tempting to make the connection between the apparent loss of function of a larval sensory gene and the exploitation of a new hostplant, as research on the silkmoth (*B. mori*) has shown that knockout of a single chemoreceptor is sufficient to turn the naturally specialist larvae into generalists (Zhang et al., 2019); however, it will take further study to establish whether or not a similar situation is occurring in *M. sexta.*

Conclusions

We report that *M. sexta* from Arizona, Kansas, and North Carolina, genetically differentiable. The confirmation of regional differentiation is a valuable addition to research in both laboratory and field studies, as well as to development of management plans for this crop pest between regions. That said, we have yet to establish the minimum geographic scale at which *M. sexta* populations are structured (for instance how much gene flow exists between adjacent states or even counties). This finer scale sampling will be particularly important for agricultural centers on the applied side, and additional sampling will also provide more insight into the large-scale variation we report here.

At least two large inversions and a potential pseudogene differentiate Arizona from the Kansas and North Carolina populations. All of these variants are high frequency (> 0.5) within Arizona but absent in the other populations. However, the nature of our sample collection limits the inferences we can make at present. For instance, the high frequency of the variants suggests they are not deleterious, but with only a single time-point of sampling in Arizona, it is hard to say whether or not these variants have reached an equilibrium frequency. Additionally, other studies have shown inversion polymorphisms to follow geographic or environmental clines (Anderson, Hoffmann, McKechnie, Umina, & Weeks, 2005; Wellenreuther, Rosenquist, Jaksons, & Larson, 2017), but, again, with only one location in Arizona surveyed, it is impossible to say if these *M. sexta* variants also follow a cline in the region. Additional sampling of the Southwest will be required to assess how widespread the inversions are.

On the functional genetic side, controlled laboratory studies will be required to assess the fitness consequences of the inversions and segregating stop-loss in a larval antennal gene we report here. In particular, the coincidence of a novel larval hostplant in Arizona and a large-effect mutation in a gene expressed in a larval chemosensory gene are suggestive but far from conclusive. The relationship between the two will require hostplant choice experiments for larvae with and without the stop-loss mutation.

Finally, efforts are underway by multiple groups to represent the diversity of life on Earth through genomic sequencing of a representative individual from large taxonomic groups (e.g. invertebrates, Robinson et al., 2011) or geographic regions (e.g. the United Kingdom, Lawniczak, Barnes, Broad, Holland, & Lewis, 2019).These projects are crucial to cataloging the variation of genes and genomes on the planet, but, as we show, may fail to capture the full scope of variation within a species. Only by

402 resequencing multiple individuals across populations within a species can we fully understand how fixed
403 or flexible genomic features are.

404 Acknowledgements

405 This project was partially funded by NSF DDIG to AJM (DEB-1701931) and DEB #1557007 and IOS
406 #1920895 to AYK. The authors wish to acknowledge Wesley Mason and Michael Hulet and the rest of
407 the Information and Telecommunication Technology Center (ITTC) staff at the University of Kansas for
408 their support with our high-performance computing. The authors also thank the University of Kansas's
409 Entomology Endowment for additional funding support in sequencing. Thanks to Paul Hime for guidance
410 on population structure analysis. Thank you to Clyde Sorenson, Judie Bronstein, and Goggy Davidowitz,
411 for initial and continued support in investigating population genomics of *M. sexta*. Kelly Dexter and
412 Rebecca Messcher helped provide the samples from University of Florida's McGuire Center collection.
413 Thanks also to the members of Ashworth 2.02 for continued patience in the discussion of the material of
414 this manuscript.

415

416

417

418

419

420

421

422

423

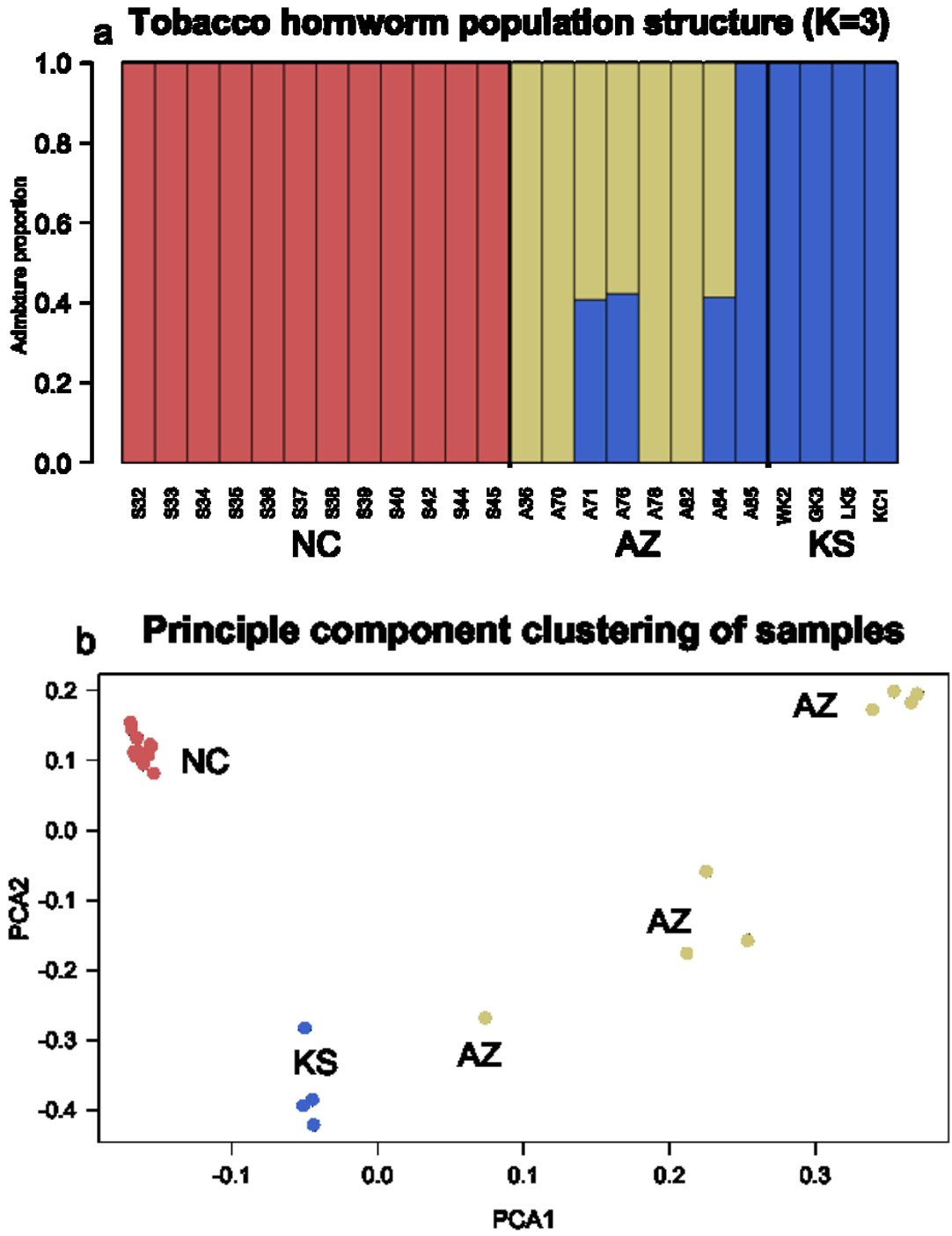

Figure 1. Population structure of *M. sexta* based on whole-genome resequencing of North Carolina (NC), Arizona (AZ), and Kansas (KS) individuals, omitting the Z chromosome. **(a)** Clustering of samples based on the most likely K (3), demonstrating that moths from different states are genetically differentiable. **(b)** Principle component analysis of the same dataset, with coloring the same as in (a). Note in both cases that Arizona individual (A85) clusters more closely with Kansas than other Arizona samples and three others (A71, A76, and A84) fall in between Kansas and the other four Arizona samples.

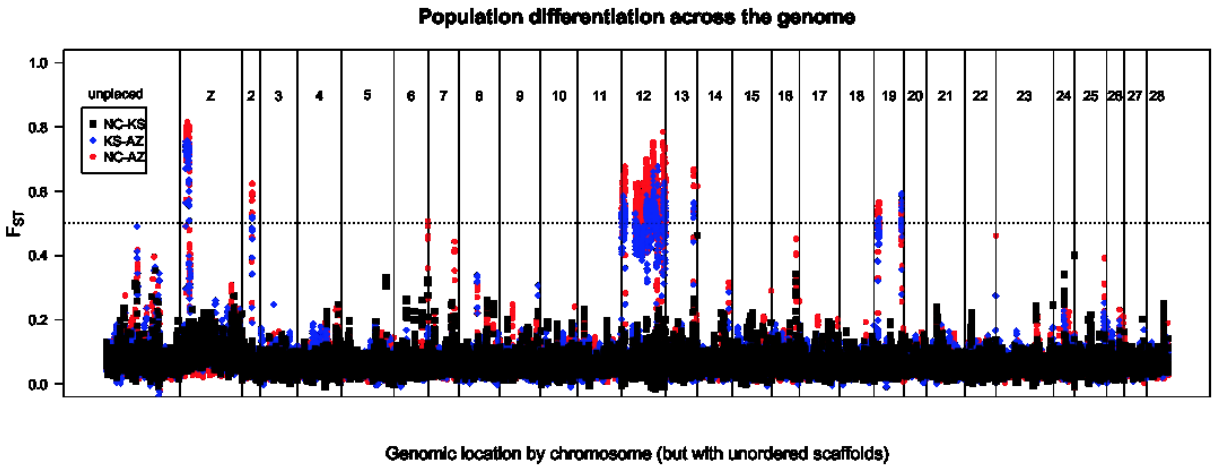

Figure 2. Pairwise differentiation (F_{ST}) across the *M. sexta* genome: North Carolina (NC) vs. Kansas (KS; black squares), Kansas vs. Arizona (AZ; blue diamonds), and North Carolina vs. Arizona (red points). Plotting is ordered by chromosomal linkage, but not ordered within a chromosome. We examined regions of differentiation greater than 0.5 (dashed line). By this metric, one peak on the Z, and one peak each on chromosomes 2, 13, and two on 19. Additionally, note the massive spike on chromosome 12, made up of numerous highly differentiated scaffolds. Note that all of these peaks come from KS vs. AZ and NC vs. AZ, indicating that Arizona's population is more distinct from the others than NC and KS are from each other.

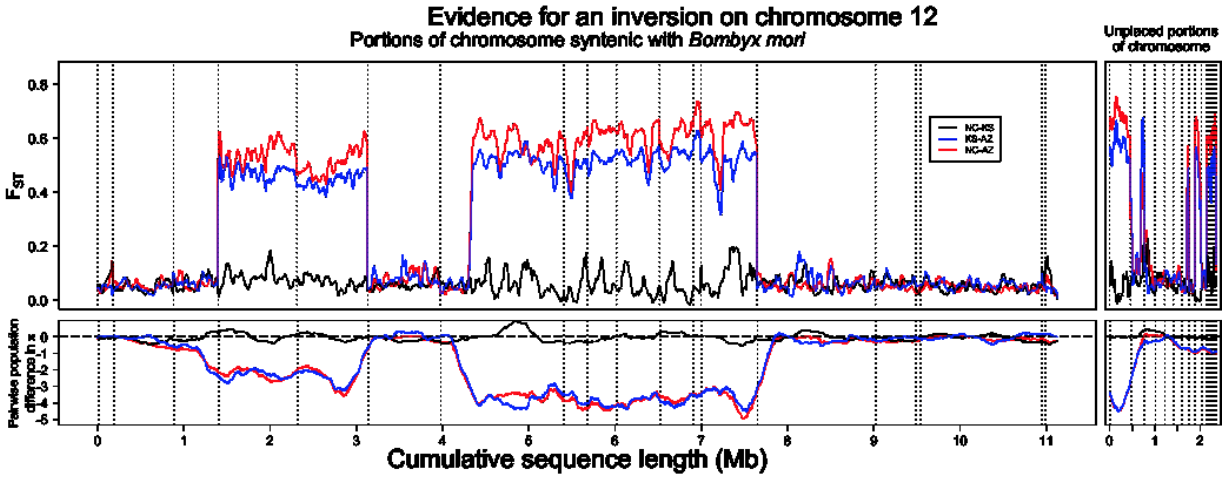

Figure 3. Investigation of differentiation across chromosome 12. **Top:** Ordering scaffolds via synteny with *Bombyx mori* suggests two regions > 1.5 Mb of elevated differentiation (F_{ST}) between *M. sexta* populations. **Bottom:** Pairwise difference in genetic diversity (π) between populations. Diversity is roughly equal (near 0 difference) except within putatively inverted regions, in which both NC and KS have few pairwise differences than AZ. In both the top and bottom, scaffold limits are denoted by dashed vertical lines. **Left:** The majority of assembled scaffolds could be placed in an order with synteny. **Right:** Some scaffolds could not be placed by this method and appear here, unordered. Note that some among these scaffolds also show scaffold-wide differentiation and differences in diversity.

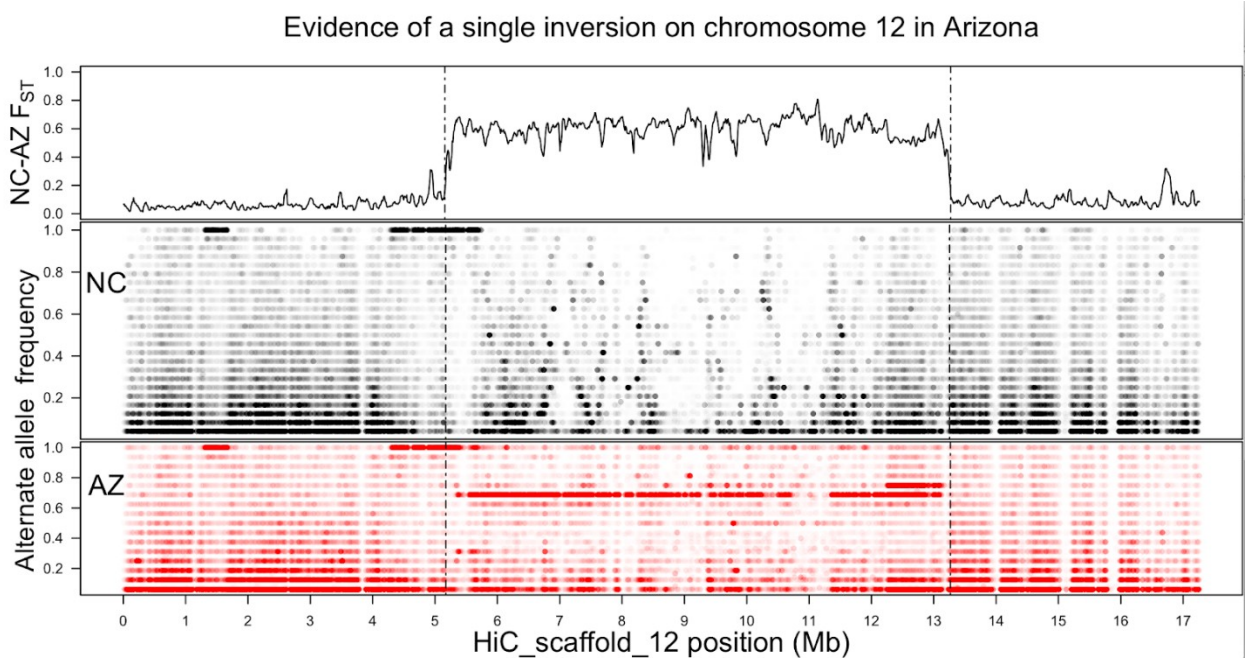

Figure 4. Further evidence of a single segregating inversion on chromosome 12 from the newly released *M. sexta* assembly (Gershman et al., 2020). **Top.** F_{ST} comparison for the North Carolina and Arizona populations shows a roughly 8 Mb region (between 5.3-13.3 Mb) of elevated differentiation between the populations, corresponding to the length of the two tracts of differentiation from Figure 3, plus some additional unplaced peaks. Vertical dashed lines delimit this differentiated region and extend below for ease of comparison between populations. **Middle and Bottom.** Frequencies of alternate alleles for called SNPs in North Carolina (middle) and Arizona (bottom). Points are semi-transparent such that opaquer regions are denser with variants. Arizona appears to hold the inverted orientation as a single allele frequency (0.6875 or 11/16) dominates the differentiated region, whereas North Carolinian allele frequencies are variable throughout. Alsoin the 1-2 Mb and 4-5 Mb regions are areas of dense SNPs fixed in both populations. These are interpreted as regions in which the lab strain used to generate the reference has diverged from the wild populations.

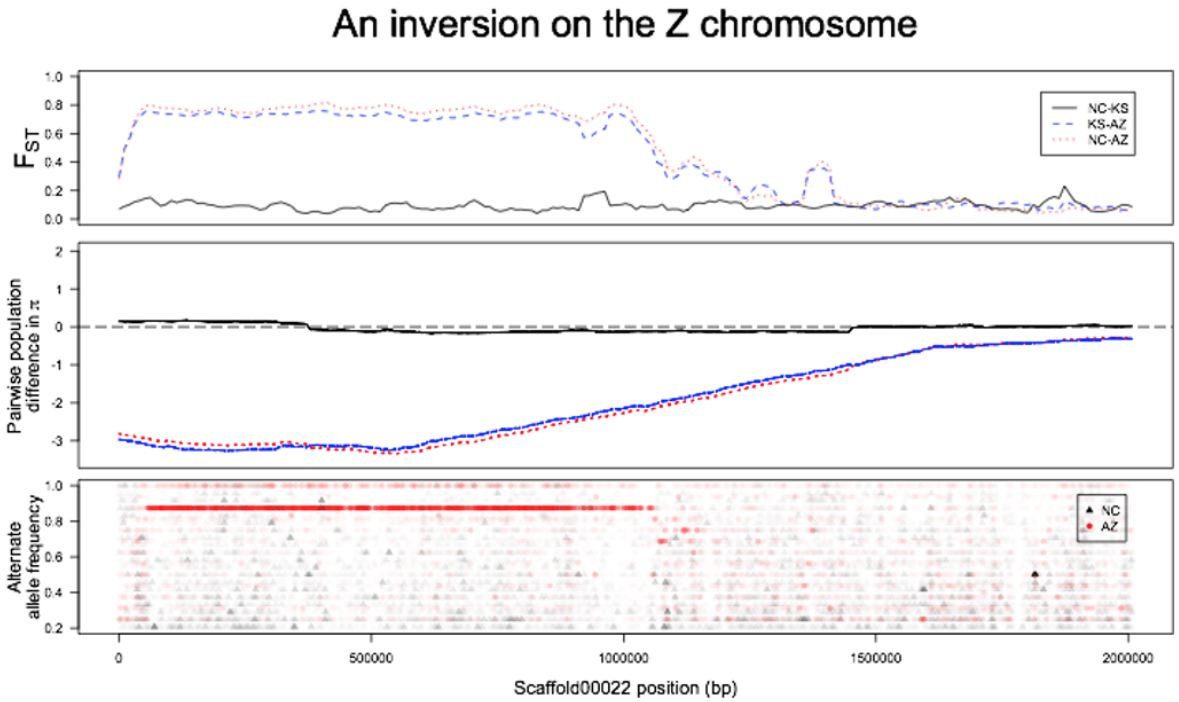

Figure 5. Evidence for an inversion on the Z chromosome. **Top.** A roughly 1 Mb portion of scaffold0002 (from 58 Kb to 1.03 Mb) shows elevated differentiation in North Carolina – Arizona and Kansas – Arizona comparisons. **Middle.** As with the chromosome 12 inversion, this region shows differences in π. **Bottom.** Like the chromosome 12 inversion, Arizona shows a long tract of shared allele frequencies in the region while North Carolina does not (note that Kansas allele frequencies are omitted due to lower sampling effort artificially stratifying frequencies).

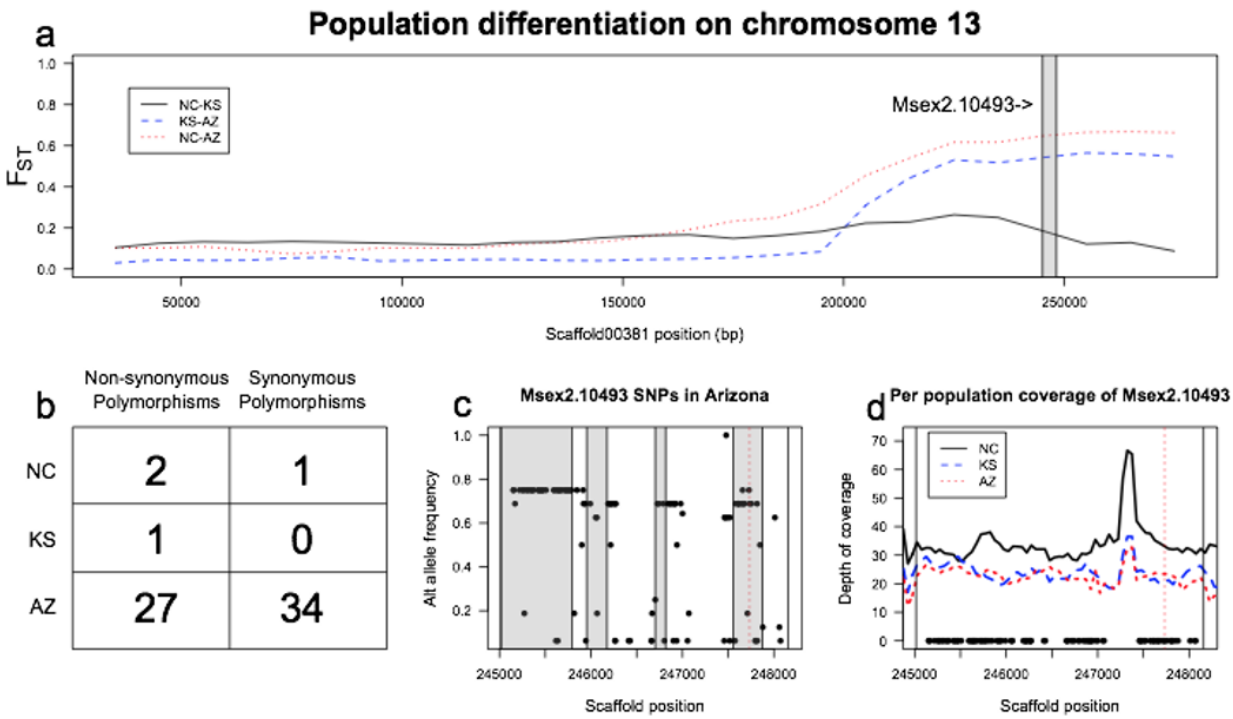

Figure 6. Population differentiation in chromosome 13. **a.** Finer scale plotting of F_{ST} of the differentiated scaffold on chromosome 13 shows that differentiation occurs beyond 200 Kb and, as with the inversions above, it appears that Arizona is markedly different from North Carolina and Kansas. A single annotated gene, Msex2.10493, lies in this region (grey box). **b.** Annotation of SNPs in the coding region reveals a stark difference in variation between populations. Arizona moths hold more than ten times the non-synonymous polymorphisms than those in the other two states. **c.** Frequencies of the SNPs in Arizona. Grey boxes represent exons and the dashed vertical line marks an annotated premature stop codon. **d.** Depth of coverage of the gene in question does not vary between populations in a way that suggests differences in alignment between populations can explain the differences in called SNPs.

	Antenna		Head	Muscle	Midgut	Malpighian tubes	Fat	Gonad	
Larva									
Pupa								M	F
Adult	M	F						M	F

Table 1. Tissue-specific RNA sequencing available for each life-stage of *M. sexta*. This dataset forms the basis for tissue specificity labeling annotation of genes in this study. Grey cells indicate presence, empty cells denote absence of data. If sex specific data are available, it is noted with an M (for male) and F (female) in the relevant cell. For larvae, few tissues were sampled at multiple instars, so no further developmental division was made.

498

499

F_{ST} 0x, 4x	NC	KS
KS	0.064, 0.054	
AZ	0.090, 0.081	0.041, 0.044

500

501 **Table 2.** Genome-wide point estimates (excluding the Z chromosome) for differentiation between each
502 pairwise combination of the three populations studied here. In each comparison differentiation at sites
503 exposed to selection (i.e. zero-fold degenerate sites, 0x) and putatively neutral sites (four-fold
504 degenerate sites, 4x) were estimated separately.

505

506 # References

507 Anderson, A. R., Hoffmann, A. A., McKechnie, S. W., Umina, P. A., & Weeks, A. R. (2005). The latitudinal
508 cline in the In (3R) Payne inversion polymorphism has shifted in the last 20 years in Australian
509 Drosophila melanogaster populations. *Molecular Ecology*, *14*(3), 851–858.

510 Banister, M. J., & White, R. H. (1987). Pigment migration in the compound eye of Manduca sexta: effects
511 of light, nitrogen and carbon dioxide. *Journal of Insect Physiology*, *33*(10), 733–743.

512 Boer, G. de, & Hanson, F. E. (1987). Differentiation of roles of chemosensory organs in food
513 discrimination among host and non-host plants by larvae of the tobacco hornworm, Manduca
514 sexta. *Physiological Entomology*, *12*(4), 387–398.

515 Cao, X., & Jiang, H. (2017). An analysis of 67 RNA-seq datasets from various tissues at different stages of
516 a model insect, *Manduca sexta*. *BMC Genomics*, *18*(1). doi: 10.1186/s12864-017-4147-y

517 Charlesworth, B., Coyne, J. A., & Barton, N. H. (1987). The relative rates of evolution of sex
518 chromosomes and autosomes. *The American Naturalist*, *130*(1), 113–146.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w 1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. doi: 10.4161/fly.19695

Contreras, H. L., Goyret, J., von Arx, M., Pierce, C. T., Bronstein, J. L., Raguso, R. A., & Davidowitz, G. (2013). The effect of ambient humidity on the foraging behavior of the hawkmoth Manduca sexta. *Journal of Comparative Physiology A*, *199*(11), 1053–1063.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, *14*(8), 2611–2620.

Ferguson, D. (1991). An essay on the long-range dispersal and biogeography of Lepidoptera, with special reference to the Lepidoptera of Bermuda. *Memoirs of the Entomological Society of Canada*, *158*, 67–79.

Gershman, A., Romer, T. G., Fan, Y., Razaghi, R., Smith, W. A., & Timp, W. (2020). De novo genome assembly of the Tobacco Hornworm moth (Manduca sexta). *BioRxiv*, 2020.08.29.270983. doi: 10.1101/2020.08.29.270983

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., Di Palma, F., & Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, *26*(9), 1145–1151.

Haber, W. A., & Frankie, G. W. (1989). A tropical hawkmoth community: Costa Rican dry forest Sphingidae. *Biotropica*, 155–172.

Höglund, G., & Struwe, G. (1970). Pigment migration and spectral sensitivity in the compound eye of moths. *Zeitschrift Für Vergleichende Physiologie*, *67*(3), 229–237.

543 Janzen, D. H. (1984). *Two ways to be a tropical big moth: Santa Rosa saturniids and sphingids*.

544 Kanost, M. R., Arrese, E. L., Cao, X., Chen, Y.-R., Chellapilla, S., Goldsmith, M., … Jiang, H. (2016).
545 Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth,
546 *Manduca sexta*. *Insect Biochemistry and Molecular Biology*, *76*, 118–147.

547 Kanost, M. R., Jiang, H., & Yu, X. Q. (2004). Innate immune responses of a lepidopteran insect, *Manduca*
548 *sexta*. *Immunological Reviews*, *198*, 97–105. doi: 10.1111/j.0105-2896.2004.0121.x

549 Kawahara, A. Y., Breinholt, J. W., Ponce, F. V., Haxaire, J., Xiao, L., Lamarre, G. P. A. A., … Kitching, I. J.
550 (2013). Evolution of Manduca sexta hornworms and relatives: Biogeographical analysis reveals
551 an ancestral diversification in Central America. *Molecular Phylogenetics and Evolution*, *68*(3),
552 381–386. doi: 10.1016/j.ympev.2013.04.017

553 Kawamoto, M., Jouraku, A., Toyoda, A., Yokoi, K., Minakuchi, Y., Katsuma, S., … Shimada, T. (2019). High-
554 quality genome assembly of the silkworm, Bombyx mori. *Insect Biochemistry and Molecular*
555 *Biology*, *107*, 53–62.

556 Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*,
557 *173*(1), 419–434.

558 Knief, U., Hemmrich-Stanisak, G., Wittig, M., Franke, A., Griffith, S. C., Kempenaers, B., & Forstmeier, W.
559 (2016). Fitness consequences of polymorphic inversions in the zebra finch genome. *Genome*
560 *Biology*, *17*(1), 199.

561 Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: a program
562 for identifying clustering modes and packaging population structure inferences across K.
563 *Molecular Ecology Resources*, *15*(5), 1179–1191.

564 Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation
565 Sequencing Data. *BMC Bioinformatics*, *15*(1). doi: 10.1186/s12859-014-0356-4

Kryuchkova-Mostacci, N., & Robinson-Rechavi, M. (2017). A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, *18*(2), 205–214. doi: 10.1093/bib/bbw008

Lande, R. (1984). The expected fixation rate of chromosomal inversions. *Evolution*, 743–752.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. doi: 10.1038/nmeth.1923

Lawniczak, M., Barnes, I., Broad, G., Holland, P., & Lewis, O. (2019). *The Darwin Tree of Life Project: Ambitions for insect genomics*. Presented at the Entomology 2019.

Lee, Y. W., Fishman, L., Kelly, J. K., & Willis, J. H. (2016). A segregating inversion generates fitness variation in yellow monkeyflower (Mimulus guttatus). *Genetics*, *202*(4), 1473–1484.

Lyons, J. I., Pierce, A. A., Barribeau, S. M., Sternberg, E. D., Mongue, A. J., & de Roode, J. C. (2012). Lack of genetic differentiation between monarch butterflies with divergent migration destinations. *Molecular Ecology*, *21*(14), 3433–3444. doi: 10.1111/j.1365-294X.2012.05613.x

McKenna, A. H., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … Depristo, M. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. doi: 10.1101/gr.107524.110

McNeil, J. (1995). Physiological integration of migration in Lepidoptera. *Insect Migration: Tracking Resources through Space and Time*.

Mechaber, W. L., & Hildebrand, J. G. (2000). Novel, non-solanaceous hostplant record for Manduca sexta (Lepidoptera: Sphingidae) in the Southwestern United States. *Annals of the Entomological Society of America*, *93*, 447–451. doi: 10.1603/0013-8746(2000)093[0447:NNSHRF]2.0.CO;2

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731.

588 Mongue, A. J., Hansen, M. E., Gu, L., Sorenson, C. E., & Walters, J. R. (2019). Nonfertilizing sperm in
589 Lepidoptera show little evidence for recurrent positive selection. *Molecular Ecology*, *28*(10),
590 2517–2530. doi: 10.1111/mec.15096
591 Mongue, A. J., Hansen, M. E., & Walters, J. R. (2020). ZZ Top: faster and more adaptive Z chromosome
592 evolution in two Lepidoptera. *BioRxiv*, 2020.06.13.142711. doi: 10.1101/2020.06.13.142711
593 Mongue, A. J., & Walters, J. (2017). The Z chromosome is enriched for sperm proteins in two divergent
594 species of Lepidoptera. *Genome*, *61*(4), 248–253.
595 Nardi, J. B. (1993). Modulated expression of a surface epitope on migrating germ cells of Manduca sexta
596 embryos. *Development*, *118*(3), 967–975.
597 Nei, M. (1987). *Molecular evolutionary genetics*. Columbia university press.
598 Nei, M., Niimura, Y., & Nozawa, M. (2008). The evolution of animal chemosensory receptor gene
599 repertoires: roles of chance and necessity. *Nature Reviews Genetics*, *9*(12), 951–963.
600 Prieto-Godino, L. L., Rytz, R., Bargeton, B., Abuin, L., Arguello, J. R., Dal Peraro, M., & Benton, R. (2016).
601 Olfactory receptor pseudo-pseudogenes. *Nature*, *539*(7627), 93–97.
602 R Core Team. (2017). R Core Team (2017). R: A language and environment for statistical computing. *R*
603 *Foundation for Statistical Computing, Vienna, Austria. URL Http://Www.R-Project.Org/.*, R
604 Foundation for Statistical Computing.
605 Rice, W. R. (1984). Sex chromosomes and the evolution of sexual dimorphism. *Evolution*, *38*(4), 735–
606 742. doi: 10.2307/2408385
607 Robinson, G. E., Hackett, K. J., Purcell-Miramontes, M., Brown, S. J., Evans, J. D., Goldsmith, M. R., …
608 Schneider, D. J. (2011). Creating a buzz about insect genomes. *Science*, *331*(6023), 1386.
609 Shapiro, A. M. (2006). Survival of freezing and subsequent summer eclosion by three migratory moths:
610 Manduca sexta and Hyles Lineata (Sphingidae), and Helicoverpa zea (Noctuidae). *Journal of the*
611 *Lepidopterists' Society*, *60*(2), 101–103.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693–702.

Stevenson, R., Corbo, K., Baca, L., & Le, Q. (1995). Cage size and flight speed of the tobacco hawkmoth Manduca sexta. *Journal of Experimental Biology*, *198*(8), 1665–1672.

The International Silkworm Genome Consortium. (2008). The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochemistry and Molecular Biology*, *38*(12), 1036–1045. doi: 10.1016/j.ibmb.2008.11.004

Wagner, D. L. (2005). *Caterpillars of eastern North America: a guide to identification and natural history*. Princeton University Press.

Wellenreuther, M., Rosenquist, H., Jaksons, P., & Larson, K. W. (2017). Local adaptation along an environmental cline in a species with an inversion polymorphism. *Journal of Evolutionary Biology*, *30*(6), 1068–1077.

Whittington, E., Zhao, Q., Borziak, K., Walters, J. R., & Dorus, S. (2015). Characterisation of the *Manduca sexta* sperm proteome: Genetic novelty underlying sperm composition in Lepidoptera. *Insect Biochemistry and Molecular Biology*, *62*, 183–193. doi: 10.1016/j.ibmb.2015.02.011

Wysoker, A., Tibbetts, K., & Fennell, T. (2013). Picard tools version 1.90. *Http://Picard. Sourceforge. Net (Accessed 14 December 2016)*, *107*(17), 308.

Yasukochi, Y., Tanaka-Okuyama, M., Shibata, F., Yoshido, A., Marec, F., Wu, C., … Sahara, K. (2009). Extensive conserved synteny of genes between the karyotypes of *Manduca sexta* and *Bombyx mori* revealed by BAC-FISH mapping. *PLoS ONE*, *4: e7465*(10). doi: 10.1371/journal.pone.0007465

Zhan, S., Zhang, W., Niitepõld, K., Hsu, J., Haeger, J. F., Zalucki, M. P., … Kronforst, M. R. (2014). The genetics of monarch butterfly migration and warning colouration. *Nature*, *514*(7522), 317–321.

Zhang, Z.-J., Zhang, S.-S., Niu, B.-L., Ji, D.-F., Liu, X.-J., Li, M.-W., … Tan, A.-J. (2019). A determining factor for insect feeding preference in the silkworm, Bombyx mori. *PLOS Biology*, *17*(2), e3000162. doi: 10.1371/journal.pbio.3000162