Impact of Introgression on Adaptation and Range Expansions

By

ERIN WALKER CALFEE

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Population Biology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Graham M. Coop, Chair

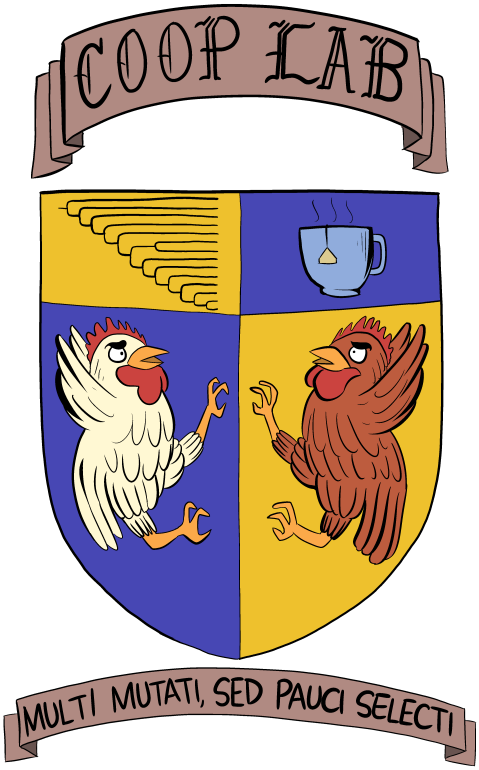
Jeffrey Ross-Ibarra

Michael Turelli

Committee in Charge

2021

To the Coop lab ... may the worst puns and best teas always prevail!



Contents

Impact of Introgression on Adaptation and Range Expansions

Abstract

Gene flow between diverged populations is often deleterious (e.g. due to genetic incompatibilities or local adaptation), but introgression can also be a source of rapid adaptation, especially to novel niches. My thesis focuses on the evolutionary outcomes of these opposing selection forces on introgressed ancestry, the repeatability of these outcomes across admixed populations, and the consequences for species niches. I have developed novel population genetics methods to detect selection in admixed populations, and applied these methods to disentangle how demography and selection have shaped the evolution and range expansions of two very different species: *scutellata*-European hybrid honey bees and highland maize.

For my first dissertation chapter, I conducted a cross-continental comparison of the outcomes of admixture and selection in *scutellata*-European honey bees. *Scutellata* honey bees from South Africa were introduced to the Americas in the 1950s, but soon escaped, and through interbreeding with European honey bees, formed a highly successful and invasive hybrid population that spread at ∼300km/year across the Americas. This is a great system to study the adaptive potential and the limitations of admixture in facilitating rapid range expansions, with natural replication in North and South America. For this research, I collected and sequenced 300+ bees from two nearly 1000 km transects, one in California and one in Argentina, to compare ancestry patterns across their genomes. I found evidence of convergent selection favoring African *scutellata* honey bee ancestry at a number of loci in the genome in North and South America. These loci are strong candidates for contributing to the high fitness and success of *scutellata*-European hybrid honey bees. Because these bees are highly defensive, their continued spread is an agricultural and public health concern. I found parallel clines in genomewide ancestry between continents at similar latitudes, despite much larger dispersal distances to reach California from the origin of the invasion, evidence that many loci across the genome are currently preventing spread to temperate zones.

For my second dissertation chapter, I analyzed the outcome of introgression between maize and its wild highland 'teosinte' relative, *mexicana*, which may have facilitated maize's range expansion from the valleys where it was domesticated up to 3000m in the mountains of Mexico. Using a

novel method that accounts for background patterns of ancestry variance and covariance between populations (e.g. due to gene flow or shared drift post-admixture), I found strong evidence for adaptive introgression from *mexicana* into maize, especially among the highest elevation populations, consistent with introgression facilitating maize's colonization of the highlands. I also found loci (including a newly identified inversion) where selection maintains steep ancestry clines across elevation. I demonstrated evidence of selection against introgression, removing *mexicana* ancestry from near domestication genes and lower recombination regions of the genome (due to linked selection). One surprising finding is that despite observations of hybrids in the field, and opportunities for gene flow from locally adapted *mexicana* that grows side-by-side with contemporary maize, I found little evidence for recent locally sourced haplotypes genomewide or at loci with high local introgression. Rather, the majority of introgression is from over 1000 generations ago, and has subsequently diverged within the maize background and been sorted by selection along an elevational cline and within individual populations. This work has broader impacts for understanding the longer term effects of introgression on range expansions and aiding in the discovery of key loci associated with high-elevation adaptations, which may be crucial for future breeding of maize, a global staple, under climate change.

Overall, this thesis adds to our knowledge of the role of introgression in range and niche expansions, and provides in-depth genomic analyses of selection and admixture in two agriculturally-important species.

Acknowledgments

First, I would like to thank Graham Coop for encouraging me to trust my ideas and find my own path as a scientist. May I grow into being as kind and thoughtful of a mentor and teacher.

To Coopons past and present, it's been a great joy and an honor to be your peer: Kristin Lee, Doc Edge, Sivan Yair, Matt Osmond, Emily Josephs, Nancy Chen, Vince Buffalo, Katie Ferris, Jeff Groh, Pavitra Muralidhar, Carl Veller, and Jeremy Berg.

I am grateful to many colleagues and friends for supporting my growth as a scientist. Thank you to Jeff Ross-Ibarra for modeling open science and brightening my time at UC Davis with your enthusiasm for Zea evolution, population genetics, and a-maize-ing puns. Thank you to Michael Turelli for treating me like a junior colleague from day one and sharing your love for good science and good wine. Thank you to Annie Schmitt and Chuck Langley for encouraging me to think broadly. To all of the staff in Evolution and Ecology, and especially Sherri Mann, thank you for your kindness and taking care of all the other things so I could focus on my research.

To all of the CPB students who met my fledgling scientific ideas with perhaps more enthusiasm then they deserved, you are the spark. A special shout-out to my PhD cohort and close friends for your laughter and support along the way: Vicky Morgan, Kelsey Lyberger, Anita To, Yige Luo, Katherine Corn, Mikaela Provost, Caitlin French, Charlotte Pickett, Matt van Avermaete, Hayley Rousek and Debbie Pattison.

Chapter 1 would not have been completed without the generosity of my collaborators: Graham Coop and Santiago Ramirez for believing in my vision for a cross-continent comparative study and welcoming me into their labs; and Marcelo Nicolás Agra and María Alejandra Palacio for helping me realize this vision with honey bee samples from Argentina. Thank you to Daniel Gates, Anne Lorant, M. Taylor Perkins, Graham Coop and Jeff Ross-Ibarra for their advice and help with Chapter 2.

This thesis would not have been completed without the love and support of my family. Thank you to my parents, Kat and Dave, and sister Alexa for all of your encouragement, and my cat Houdini for being at my side the whole way. Lastly, thank you to my husband Joshua Foster for reminding me there are many mountains to climb (including my PhD) and I can summit any of them if I just keep going up.

CHAPTER 1

Selection and hybridization shaped the rapid spread of African honey bee ancestry in the Americas

Erin Calfee[1,2], Marcelo Nicolás Agra[3], María Alejandra Palacio[3,4], Santiago R. Ramírez[1,2], and Graham Coop[1,2]

[1] Center for Population Biology, University of California, Davis, United States of America

[2] Department of Evolution and Ecology, University of California, Davis, United States of America

[3] Instituto Nacional de Tecnología Agropecuaria (INTA), Balcarce, Argentina

[4] Facultad de Ciencias Agrarias, Universidad de Mar del Plata, Balcarce, Argentina

Abstract

Recent biological invasions offer 'natural' laboratories to understand the genetics and ecology of adaptation, hybridization, and range limits. One of the most impressive and well-documented biological invasions of the 20th century began in 1957 when *Apis mellifera scutellata* honey bees swarmed out of managed experimental colonies in Brazil. This newly-imported subspecies, native to southern and eastern Africa, both hybridized with and out-competed previously-introduced European honey bee subspecies. Populations of *scutellata*-European hybrid honey bees rapidly expanded and spread across much of the Americas in less than 50 years. We use broad geographic sampling and whole genome sequencing of over 300 bees to map the distribution of *scutellata* ancestry where the northern and southern invasions have presently stalled, forming replicated hybrid zones with European bee populations in California and Argentina. California is much farther from Brazil, yet these hybrid zones occur at very similar latitudes, consistent with the invasion having reached a climate barrier. At these range limits, we observe genome-wide clines for *scutellata* ancestry, and parallel clines for wing length that span hundreds of kilometers, supporting a smooth transition

from climates favoring *scutellata*-European hybrid bees to climates where they cannot survive winter. We find no large effect loci maintaining exceptionally steep ancestry transitions. Instead, we find most individual loci have concordant ancestry clines across South America, with a build-up of somewhat steeper clines in regions of the genome with low recombination rates, consistent with many loci of small effect contributing to climate-associated fitness trade-offs. Additionally, we find no substantial reductions in genetic diversity associated with rapid expansions nor complete dropout of *scutellata* ancestry at any individual loci on either continent, which suggests that the competitive fitness advantage of *scutellata* ancestry at lower latitudes has a polygenic basis and that *scutellata*-European hybrid bees maintained large population sizes during their invasion. To test for parallel selection across continents, we develop a null model that accounts for drift in ancestry frequencies during the rapid expansion. We identify several peaks within a larger genomic region where selection has pushed *scutellata* ancestry to high frequency hundreds of kilometers past the present cline centers in both North and South America and that may underlie high-fitness traits driving the invasion.

Author Summary

Crop pollination around the world relies on native and introduced honey bee populations, which vary in their behaviors and climatic ranges. *Scutellata*-European hybrid honey bees (also known as 'Africanized' honey bees) have been some of the most ecologically successful; originating in a 1950s experimental breeding program in Brazil, they rapidly came to dominate across most of the Americas. As a recent genetic mixture of multiple imported *Apis mellifera* subspecies, *scutellata*-European hybrid honey bees have a patchwork of ancestry across their genomes, which we leverage to identify loci with an excess of *scutellata* or European ancestry due to selection. We additionally use the natural replication in this invasion to compare outcomes between North and South America (California and Argentina). We identify several genomic regions with exceptionally high *scutellata* ancestry across continents and that may underlie favored *scutellata*-European hybrid honey bee traits (e.g. *Varroa* mite resistance). We find evidence that a climatic barrier has dramatically slowed the invasion at similar latitudes on both continents. At the current range limits, *scutellata* ancestry decreases over hundreds of kilometers, creating many bee populations with intermediate

scutellata ancestry proportions that can be used to map the genetic basis of segregating traits (here, wing length) and call into question the biological basis for binary 'Africanized' vs. European bee classifications.

Introduction

Diverging lineages often spread back into secondary contact before reproductive isolation is complete, and so can hybridize. In hybrid zones, multiple generations of admixture and backcrossing create a natural experiment in which genetic variation is 'tested' in novel ecological and genomic contexts. The mosaic of ancestries in hybrid zones has allowed researchers to uncover the genetic loci associated with reproductive barriers (e.g. [**1**, **2**, **3**]) and to identify rapidly introgressing high-fitness alleles (e.g. [**4**, **5**, **6**, **7**]). One promising way forward is to compare ancestry patterns across multiple young hybrid zones and test how repeatable the outcome of hybridization is across these evolutionary replicates.

In this study, we use this powerful comparative framework to better understand the genomic basis of fitness and range limits of *scutellata*-European hybrid honey bees, with replicate routes of invasion into North and South America. The range of the western honey bee (*Apis mellifera*) has expanded from Africa, Europe, and western Asia [**8**] across much of the globe, assisted by colonialism and the ecological diversity of honey bee subspecies [**9**]. While the Americas have many species of native bees and a long cultural history of beekeeping with honey-producing stingless bees (Meliponini), colonists as early as the 1600s imported European honey bee subspecies for their own apiculture and agriculture uses [**10**], setting off the first honey bee invasion of the Americas [**11**]. Through a combination of human-assisted migration and swarming, European honey bees spread across the continent and founded feral populations [**10**]. Then in 1957, swarms from a newly-imported honey bee subspecies from southern and eastern Africa, *Apis mellifera scutellata*, escaped from an experimental breeding program in Brazil and rapidly dispersed. Widely successful, *scutellata* honey bees both out-competed and hybridized with European-ancestry populations, creating a rapidly advancing *scutellata*-European admixed population that expanded north and south across the Americas at 300-500 km/year [**12**].

Colonies of *scutellata*-European hybrids are likely to respond more strongly to disturbances than colonies with European ancestry (measured as number of stings per minute, reduced time to sting, and longer pursuit distances [**13**, **14**, **15**]). The spread of these more defensive bees (sensationalized as 'killer bees', see critiques [**16**, **17**, **18**, **19**]) have created new challenges for beekeepers and public health [**16**, **20**]

Control efforts have been largely unsuccessful in slowing the invasion or preventing the spread of *scutellata* ancestry into commercial colonies [**12**,**16**]. However, even without intervention, *scutellata* ancestry is unlikely to outcompete European ancestry in the coldest regions of the Americas because *scutellata*-European hybrid honey bees from the neotropics have low overwinter survival in climates where European bees thrive [**21**, **22**]. Models based on winter temperatures and the physiological cost of thermo-regulation predict northern range limits for the invasion that vary from the Central Californian Coast [**23**, **24**] up to the border with Canada [**25**]. Thus, the expected impact of *scutellata* ancestry on agriculture and queen bee production in the United States is still poorly defined. Broad surveys show that *scutellata*-like mtDNA and phenotypes are common in northern Argentina and the southern US, and drop off towards more temperate latitudes, indicating that the rapid spread of these traits has dramatically slowed, if not stopped, on both continents [**23**, **26**, **27**, **28**, **29**, **30**, **31**]. However, we lack a genome-wide view of the range limits of *scutellata* ancestry and do not know whether individual high-fitness alleles have already introgressed into higher latitudes.

Previous genomic work on the invasion has shown that *scutellata*-European hybrid honey bees are a genetic mixture of three major genetic groups: A from Africa, C from eastern Europe and M from western Europe [**32**, **33**, **34**, **35**, **36**, **37**]. Historical sources indicate that the A ancestry is from *A. m. scutellata* [**38**, **39**], while both M and C ancestries are mixtures of multiple subspecies imported from Europe, e.g. *A. m. ligustica* (C), *A. m. carnica* (C), *A. m. mellifera* (M), and *A. m. iberiensis* (M) [**10**]. Many names have been used previously to refer to *scutellata*-European hybrids in the literature, including 'African honey bees', 'African hybrid honey bees', or 'Africanized honey bees', and the ambiguous acronym 'AHB', with these names being used to describe bees identified as having *scutellata* ancestry on the basis of behavior, morphology, mtDNA, or a range of *scutellata* autosomal ancestry. Given the wide range of *scutellata*-European ancestry that we find in this study, and that *A. m. scutellata* is only one of at least 10 ecologically diverse *Apis mellifera* subspecies

native to Africa [**38**], we will simply use the label *scutellata*-European hybrids for individuals whose autosomal genome is comprised of a mixture of these ancestries.

While the key genes remain unknown, *scutellata*-European hybrid honey bees diverge from European-ancestry bees on a number of traits that may have given them a selective advantage during the invasion: they have higher reproductive rates (including faster development times, proportionally higher investment in drone production and more frequent swarming to found new colonies [**12**]), they have higher tolerances to several common pesticides [**40**], and they prove less susceptible to *Varroa* mites, a major parasite [**41**, **42**, **43**, **44**, **45**]. Population monitoring studies show that *Varroa* mites are a strong selective force in the wild and that mite infestations in the 1990s likely contributed to the rapid genetic turnover of feral nest sites from European to *scutellata*-European hybrid colonies in Arizona and Texas [**28**, **29**, **36**]. European ancestry may have also contributed to the success of the invasion; a recent study of *scutellata*-European hybrid bees in Brazil revealed some European alleles at exceptionally high frequency, but this work was underpowered to detect high-fitness *scutellata* alleles due to elevated genome-wide *scutellata* ancestry (84%) in the Brazilian population [**37**].

There are also a number of candidate traits that distinguish *scutellata*-European hybrid honey bees from European bees and plausibly contribute to a climate-based range limit for the invasion. Smaller bodies [**46**] and higher metabolic rates [**25**], for example, could give honey bees with high *scutellata* ancestry a competitive advantage in the tropics but come at a cost in cooler climates [**24**]. In addition to physiological traits, heritable behaviors may also contribute to fitness trade-offs: *scutellata*-European hybrid bees from Venezuela to Arizona preferentially forage for protein-rich pollen (vs. nectar), which supports rapid brood production, but risks insufficient honey stores to thermo-regulate over winter [**24**, **47**, **48**].

Other traits associated with *scutellata* ancestry are of central importance to beekeepers, but their role in the invasion is less clear. Stronger colony-defense behaviors have been reported across much of the range of *scutellata*-European hybrid honey bees [**16**, **39**] (with some local exceptions, see [**49**, **50**]). The fitness consequence of these behaviors will depend on the costs of both predation and defense. Similarly, more frequent absconding (leaving a nest site to find another) is undesirable in managed apiaries, but may be adaptive in some environments, e.g. to escape predators or local

resource shortages [**51**]. Selection for these traits is likely to vary across the range of *scutellata*-European hybrid honey bees, depending on the natural and human-mediated environment.

Here we conduct the first comparative study of the *scutellata*-European hybrid honey bee invasion in North and South America. First, we use broad geographic sampling and whole genome sequencing to map the present-day ancestry clines on both continents, and assess the evidence for a climatic barrier preventing the further spread of *scutellata* ancestry. Next, we use genetic diversity within *scutellata* ancestry to study the shared bottleneck within and amongst populations due to the rapid expansion during the invasion. Finally, we develop a null model that includes recent drift and use this model to test for outlier loci that may underlie high-fitness *scutellata*-European hybrid honey bee traits and climatic barriers.

Results

To survey the current geographic distribution of *scutellata* ancestry in the Americas, we sampled and sequenced freely foraging honey bees across two latitudinal transects, one in California and one in Argentina, formed from the northern and southern routes of invasion out of Brazil (Fig 1.1). We generated individual low-coverage whole-genome sequence data for 278 bees, and added to this data set 35 recently published high-coverage bee genomes from 6 additional California populations sampled 3-4 years prior [**34**]. We inferred genome-wide ancestry proportions for each individual using NGSAdmix [**52**] assuming a model of 3 mixing populations, which clearly map to the *scutellata* (A), eastern European (C), and western European (M) reference panels (Figs 1.1 and 1.9). We leveraged the fact that admixed *scutellata*-European honey bee populations were formed through a recent mixture of known genetic groups to infer the mosaic of A, C and M ancestry tracts across the genome of each bee. For each population, we applied a hidden Markov model that jointly infers the maximum likelihood single-pulse approximation for the generations since mixture and posterior probabilities for local ancestry state, based on read counts from low-coverage sequence data (ancestry_hmm [**53**]). The average local ancestry estimates within individuals agree closely with the NGSAdmix genome-wide ancestry estimates (Fig 1.10, Pearson's $r \geq 0.985$), with the HMM estimating slightly higher minor ancestry for low admixture proportions, likely as a result of some miscalled blocks. Time estimates vary by population, with a median of 47.6 generations in

the 62 years since the initial dispersal of *scutellata* queen bees out of São Paulo (see Fig 1.12 for all time estimates). In this section, we first focus on the distribution of genome-wide 'global' ancestry patterns across the two clines, which we will later compare to the variation in local ancestry at individual loci.

We observe wide hybrid zones mirrored in North and South America. In Argentina, we find the cline in ancestry spans nearly 900km, from 77% *scutellata* (A) ancestry in the north to less than 5% to the south in Buenos Aires Province. The current geographic range of A ancestry in South America is broadly consistent with prior studies using a smaller number of genetic markers (e.g. [**26**, **27**, **35**, **59**]), though the geographic and genetic resolution of these studies is too limited for detailed comparison. In North America, we find that honey bees in California have up to 42% A ancestry in the south, tapering down to approximately 0% in Davis, our northernmost sampling site. In comparison, earlier extrapolations based on mitochondrial surveys may have somewhat overestimated genome-wide A ancestry in California (e.g. 65% of foraging bees in San Diego County [**30**] and 17% in Monterey County [**31**] carry A mtDNA haplotypes). We also find excess A-like mtDNA diversity in California. While this finding is potentially consistent with *scutellata* maternal lines being favored during the expansion into Southern California, this pattern is not strongly replicated in South America and even in North America, A mitochondria do not appear to have introgressed far past the northern range limit for nuclear A ancestry (Fig 1.33).

Alongside our genomic cline, we find a corresponding phenotypic cline in worker fore wing size: closer to the equator, sampled bees have increasing A ancestry and shorter wings (Fig 1.2). By fitting a linear model to predict wing length from genome-wide ancestry, we find that A ancestry can explain a difference of -0.72mm, approximately an 8% reduction in wing length ($P = 3.65 \times 10^{-23}, R^2 = 0.31, n = 269$; see Fig 1.14). We tested for a main effect and an interaction term for the South American continent, and found no significant differences in wing length ($P = 0.81$) or its association with ancestry ($P = 0.86$) between the two clines. Thus, in contrast to the rise of dispersal-enhancing traits in other recent invasions (e.g. [**60**, **61**, **62**, **63**]), we see no evidence of a bias for longer wings at larger dispersal distances (California). Genetic crosses have shown that wing length differences between ancestries have a genetic basis [**15**] and the wing length patterns we observe here are consistent with expectations of an additive polygenic cline based on

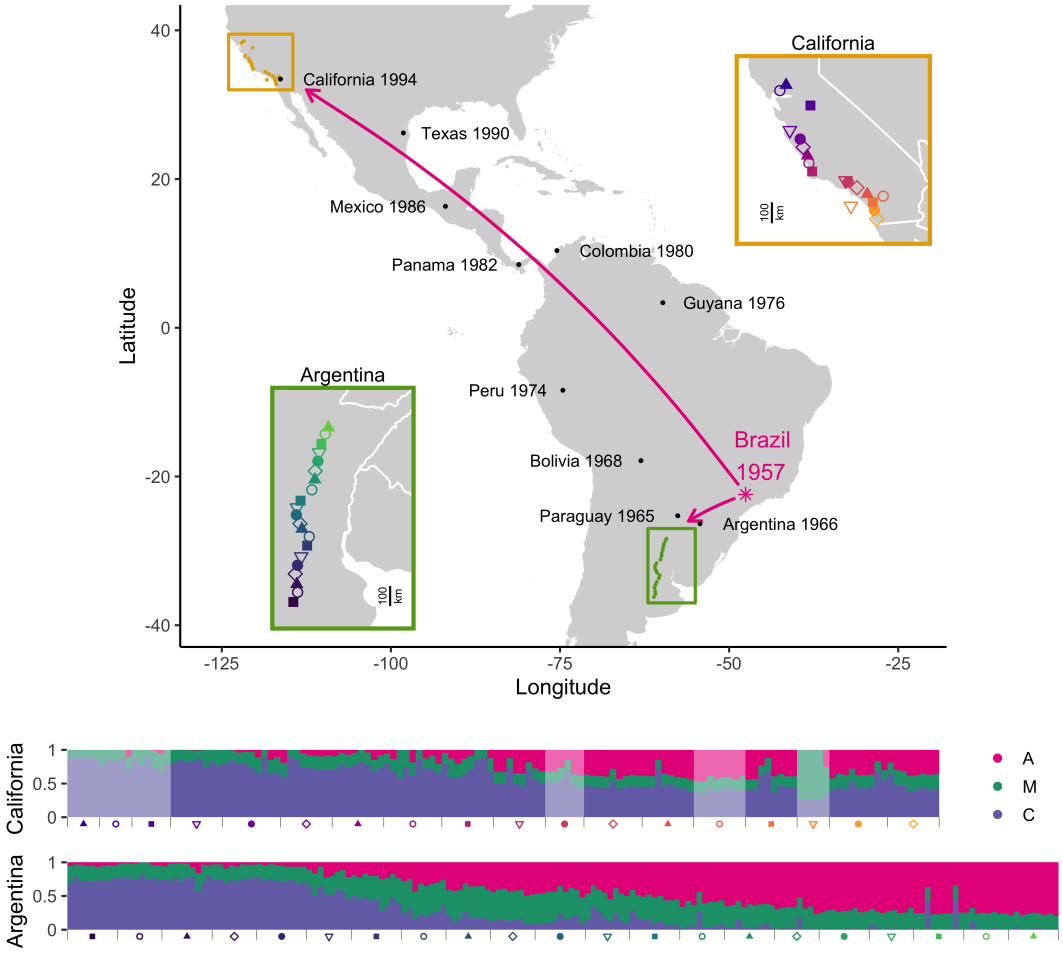


FIGURE 1.1. **Spread of *scutellata* ancestry in the Americas.** Map of hybrid zones in California and Argentina, with cartoon arrows depicting the two routes of *scutellata*-European hybrid honey bee invasion out of Rio Claro, São Paulo, Brazil. Dates of first occurrence along the routes of invasion are from [**12**] [**54**] and [**55**], with approximate GPS locations extracted from google maps. Insets zoom in on each hybrid zone to show the mean GPS coordinates for each sampled population. Sampling spanned 646km in California and 878km in Argentina in the north-south direction. Genome-wide *scutellata* (A), eastern European (C), and western European (M) ancestry inferred using NGSAdmix for each bee are shown in a bar chart at the bottom, where each vertical bar is one bee and colors indicate proportion ancestry. Populations are arranged by latitude, with samples closest to Brazil on the right. Light fading indicates that a bee comes from the previously published California data set [**34**] and was collected in the field 3-4 years prior to the bees from this study. These earlier California samples include one island population, Avalon (Catalina Island), indicated by a yellow triangle. Bees from Avalon have majority M ancestry, in contrast to all mainland California bees which have predominantly A and C ancestry. The underlying maps were created by plotting geographic data from the CIA World DataBank II [**56**] in R [**57**] using ggplot [**58**].

genome-wide ancestry alone (Fig 1.2). However, these phenotypic clines could alternatively be caused purely by developmental plasticity or sorting of within-ancestry genetic variation along a latitudinal gradient. Preliminary evidence that other factors may contribute to the wing length clines observed here comes from a 1991 survey showing that wing length was positively correlated with latitude in California's feral bee populations before the reported arrival of *scutellata*-European hybrid honey bees [**64**]. From field-based sampling alone, it remains unclear what portion of the observed phenotypic clines are ancestry-driven. We performed admixture mapping to test for genetic loci underlying ancestry-associated differences in wing length and did not identify any loci meeting genome-wide significance (Fig 1.16).

Our genomic results indicate that the geographic distribution of *scutellata* ancestry is presently constrained by climatic barriers, not dispersal. Historical records document an initial rapid spread of *scutellata*-European hybrid honey bees from their point of origin in Rio Claro, São Paulo, Brazil, followed by the formation of seemingly stable hybrid zones at similar latitudes in North and South America. Yet to reach this same latitude, northern-spreading bees had to travel more than five times the distance as southern-spreading *scutellata*-European hybrid honey bee populations.

To more precisely infer the current shape and position of the two hybrid zones, we fit a classic logistic cline model to inferred genome-wide individual *scutellata* ancestry proportions [**65**, **66**, **67**]:

$$A_i = \frac{M}{1 + e^{-b(x_i - c)}}, \tag{1.1}$$

where A_i is the genome-wide *scutellata* ancestry proportion inferred for the i^{th} individual bee, x_i is their latitude, M is the asymptotic maximum *scutellata* ancestry approaching the equator, which is set at 0.84 (i.e. frequency in Brazil [**37**]), c is the cline center, and $w = |4/b|$ is the cline width (i.e. the inverse of the steepest gradient at the center of the cline).

Each degree latitude corresponds to approximately 111km and presents a natural way to compare cline position and shape between the two zones. We fit this model in R using non-linear least squares (although maximum likelihood or Bayesian estimation are generally preferred when the errors can be fully parameterized, here least squares allows for unknown drift variance in addition to binomial sampling variance). We find that the two hybrid zones have strikingly similar

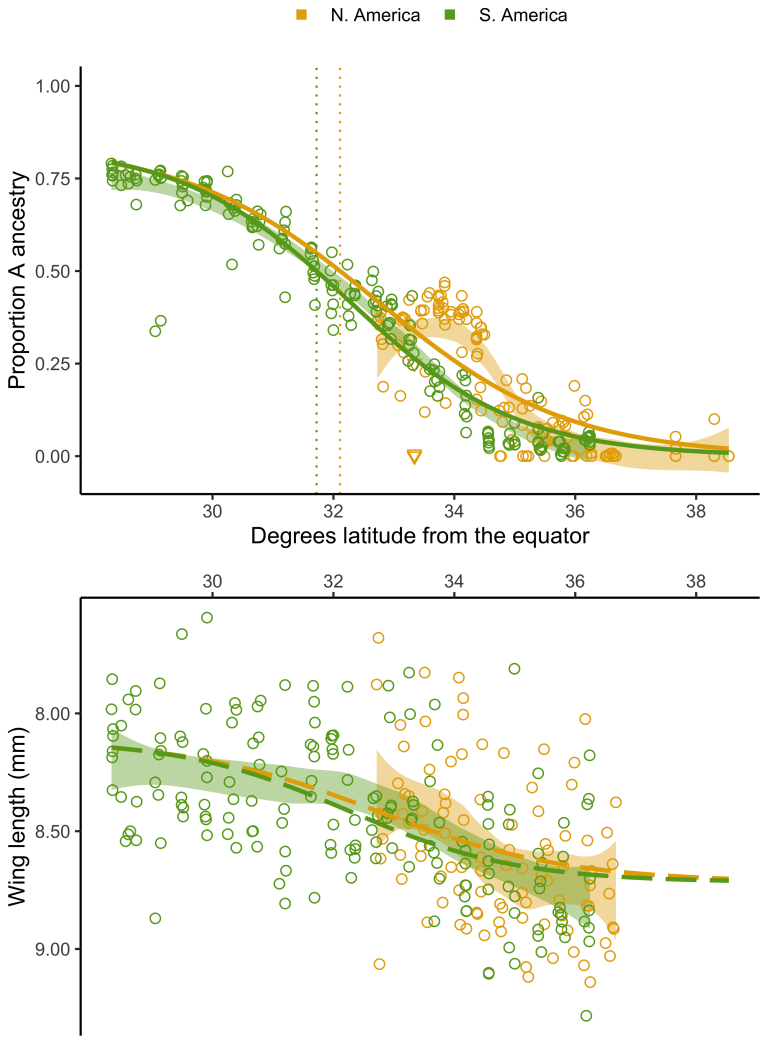


FIGURE 1.2. **Clines across latitude.** Genome-wide ancestry estimates (top) and fore wing lengths (bottom) for individual bees, plotted across absolute latitude and colored by continent. Shading indicates the 95% confidence intervals for loess curves of the raw data. We also overlay several model-fitted clines: In the top panel, solid curves show the North and South American logistic cline fits for ancestry predicted by latitude, with dotted vertical lines marking the latitude at which bees have predicted 50% *scutellata* (A) ancestry, based on these curves. Samples from Avalon are displayed as orange triangles; Catalina Island has a distinct ancestry composition from mainland California populations and low A ancestry for its latitude. In the bottom panel, dashed curves show the expected phenotypic cline if wing lengths were fully determined by the clines in ancestry depicted in the top panel. To get these predicted wing lengths, we used the mean ancestry cline as input to the best-fit linear model between ancestry and wing length. Note that the y-axis for wing lengths is reversed (smaller wings are higher) to simplify visual comparisons between the top and bottom panels.

positions (Fig 1.2), with cline centers that differ by less than half a degree (32.72°N vs. 32.26°S), and no statistically significant difference in cline steepness. To better understand the mechanisms underlying this parallelism between continents, we tested four possible explanatory climate variables to see if we could identify a better predictor for *scutellata* ancestry across our two zones than latitude: Mean annual temperature (°C), mean temperature of the coldest quarter (°C), minimum temperature of the coldest month (°C), and mean annual precipitation (cm) (downloaded from WorldClim.org [**68**]). We fit clines for both hybrid zones jointly using these four environmental variables in turn as predictors in Eq 1.1 in place of x_i, and compared these results to fits based on absolute latitude and, as a neutral dispersal model, distance from São Paulo.

We find that latitude is the best individual predictor of genome-wide global ancestry, and mean annual temperature the second-best predictor, as assessed by AIC (see Table 1.2). While latitude provides the best-fitting cline, we find it unlikely that latitude or daylight per se is the relevant selection gradient. Temperature and precipitation are closely coupled to latitude across our transect in Argentina, so nearly all of our resolution to disentangle latitude from environmental gradients comes from micro-climates within California, and for precipitation, the contrast between continents (Fig 1.13). However, we failed to identify specific environmental variables that may be driving the relationship with latitude, either because we did not include the relevant environmental variable(s) or because the climate data does not reflect the selection environment of sampled bees, e.g. due to mismatches in scale or selective habitat use by bees within a foraging range.

Despite limited resolution on the climate variables driving the latitudinal gradient, our comparative framework allows us to firmly reject a neutral model based on distance from the point of introduction in Brazil, because a single dispersal rate cannot generate predictions that simultaneously fit the clines in North and South America well (see Table 1.2).

In addition to these global ancestry estimates, we measure variation in local ancestry frequencies across the genome, which are informative about recent evolutionary history. *Scutellata* ancestry frequencies at individual loci will vary around their genome-wide mean due to finite sampling, but also evolutionary processes, including drift and selection. If two populations have shared gene flow post-admixture, at loci where one population has higher than average *scutellata* ancestry frequencies, the second population will also tend to have higher than average *scutellata* ancestry.

We capture this genetic signature in an ancestry covariance matrix, where each entry represents how much a pair of populations co-deviate in locus-specific *scutellata* ancestry away from their individual genome-wide means (Fig 1.3). We expect ancestry co-variances to build up along each route of the *scutellata*-European hybrid honey bee invasion as a result of shared drift post-admixture. Indeed, we do observe positive ancestry covariances for nearby populations within each hybrid zone. We attribute this pattern to shared demographic history, but also note that weak selection for a specific ancestry at many loci genome-wide could also generate these positive covariances. Unexpectedly, we find that populations in more temperate North and South America, i.e. at opposite ends of the expansion, have higher ancestry correlations with each other than with populations situated between them. This robust signal is a general pattern that holds true on average across chromosomes (Fig 1.21), and so isn't driven by individual outlier loci, and persists across recombination rate bins (Fig 1.22). These similar ancestry patterns in geographically distant populations are potentially consistent with a genome-wide signature of convergent selection to cooler climates or convergent selection by beekeepers at higher latitudes. Another possible explanation is recent long-distance migration (e.g. international bee exports); however, we investigated genetic covariance patterns within A, C, and M ancestries and found no clear evidence of gene flow between the high-latitude cline endpoints (see methods).

Genetic basis of the climate barrier. To identify loci that may be contributing to a climate barrier, we looked for loci with steeper than expected ancestry clines across latitude in South America. We estimated best-fitting logistic ancestry clines at ∼542k single nucleotide polymorphisms (SNPs) across the genome by re-fitting eqn. 1.1, where x_i is the population latitude and A_i is the population-mean local *scutellata* ancestry at a focal SNP, and the maximum *scutellata* ancestry M is 1. Similar cline models have been fit using likelihood methods under some simplifying assumptions about the form of the errors (e.g. [**66**, **67**]). We instead use non-linear least squares to fit cline parameters without specifying a full error model and then quantify the effects of more complex unmodeled errors (including ancestry variances and covariances) through simulation. We simulated data for 100,000 independent loci undergoing drift, which we used to estimate the expected distribution of neutral clines and calculate false-discovery rates. For each simulated locus, we independently drew a vector of population ancestry frequencies from a multivariate-normal

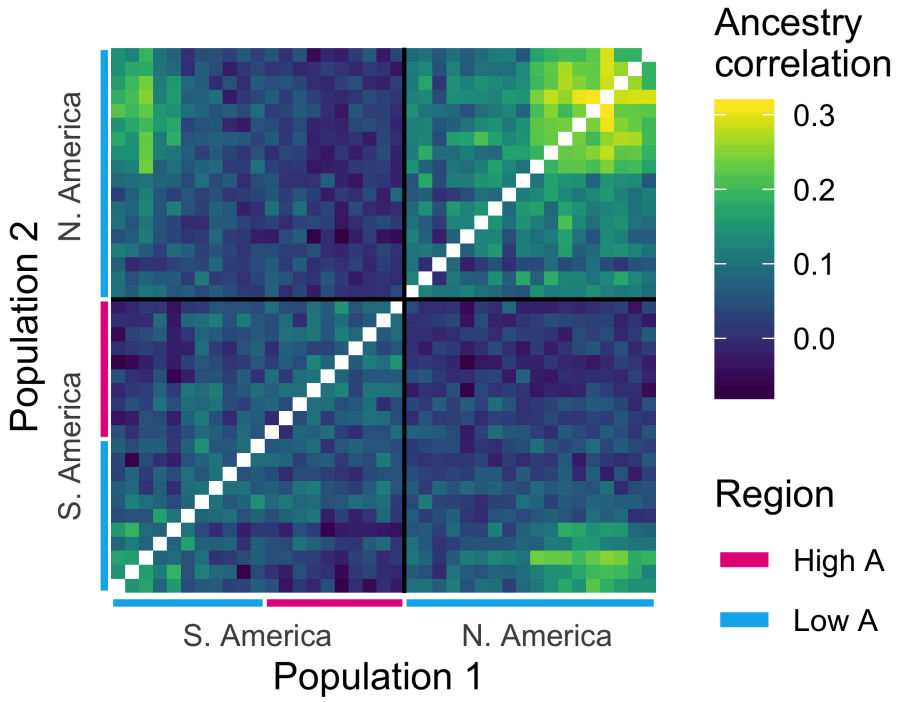


FIGURE 1.3. **Correlated ancestry across populations.** Shared drift in ancestry shown as an ancestry correlation matrix (see methods). Populations are ordered by latitude and diagonals are left blank (within-population correlations = 1). Low and high A ancestry regions of each hybrid zone are defined relative to the estimated latitude of the cline center for genomewide ancestry. About half of the sampled South American populations, and all of the North American populations, fall in the 'low A' half of their respective hybrid zones.

model of drift, A ancestry $\sim$ MVN(α, K), where α is the vector of population mean genome-wide *scutellata* (A) ancestry proportions, and the K matrix measures the expected variance and covariance in ancestry away from this mean (Fig 1.3), and is empirically calculated using all loci across the genome (see methods for additional details). We limit the analysis of clines at individual loci to South America where, unlike North America, we have samples spanning both halves of the hybrid zone to inform parameter estimates. While cline analyses can be used to identify both adaptive introgression and barriers to introgression (by analysing cline center displacement in addition to cline steepness [**69**]), here we focus on barrier loci and approach identifying positively selected loci using alternative methods that can be applied to both hybrid zones (see "Scan for ancestry-associated selection").

We find no evidence to support a simple genetic basis or environmental threshold to the climate barrier. Ancestry clines in South America are reasonably concordant across most SNPs; 95% of cline centers fall within a 1.6 degrees latitude range, with a long tail that appears to be due to adaptively introgressing loci (identified as outliers below, see Fig 1.17). We find no strongly selected individual barrier loci that exceed our 5% false-discovery-threshold for cline steepness, set by MVN simulation of background ancestry patterns. On average, individual SNP clines in South America are 960km wide (w = 8.65 degrees latitude), and the steepest cline in the genome still takes approximately 555km (w = 5 degrees latitude) to fully transition from *scutellata* to European ancestry. These wide clines, coupled with the evidence for parallel genome-wide clines in North and South America, are consistent with selection tracking smooth climate transitions over broad geographic regions rather than a discrete environmental step. Furthermore, concordance in clines across SNPs in South America suggests that many loci are associated with climate-based fitness trade-offs. Under a polygenic climate barrier, we expect locally-adapted loci to be found across the genome but steeper clines to be more commonly maintained in regions with low recombination rates. This is because selected loci create stronger barriers to gene flow when there is tight genetic linkage than when selection acts on each locus independently [**70**]. We test this theoretical prediction in South America and find enrichment for steeper clines in regions of the genome with low recombination. The empirical top 5% steepest clines in South America are found on all 16 chromosomes and are enriched in regions of the genome with low recombination. Steep clines comprise 12.7% CI_{95}[8.4%-16.5%] of loci from the lowest recombination rate quintile vs. only 3.3% CI_{95}[3.0%-3.6%] of loci from the highest recombination rate quintile. The average effect of recombination is a 50km decrease in mean cline width between the highest and lowest recombination rate quintiles (Δb = 0.028, [0.017-0.038]).

Diversity and rapid expansion. From their point of origin in Brazil, *scutellata*-European hybrid honey bees invaded much of the Americas in less than 50 years [**39**]. Such rapid expansion can lead to high rates of drift in the continually bottle-necked populations at the front of the wave of expansion, i.e. those populations sampled furthest from Brazil. To test this expectation, we calculated nucleotide diversity, π, for each sampled population (Fig 1.4). Despite much further distances traveled to the northern hybrid zone, we do not observe a more pronounced bottleneck

in California than in Argentina, suggesting that the expanding wave of *scutellata*-European hybrid honey bees maintained large population sizes (and did not experience strong 'allele surfing' [**71**]).

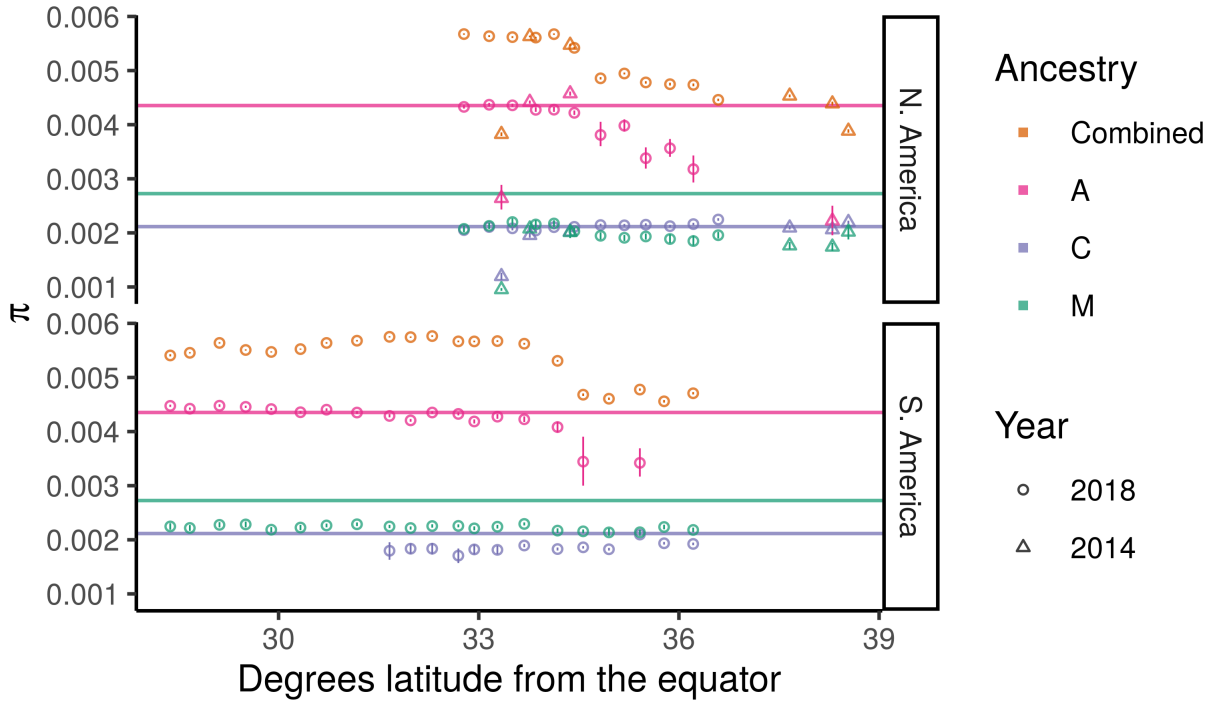


FIGURE 1.4. **Allelic diversity (π) across the hybrid zones.** For each population, we estimated allelic diversity genome-wide and within high-confidence homozygous ancestry states. Horizontal lines show the genome-wide diversity within the reference panels. Vertical lines show the 95% confidence interval for each estimate, based on a simple block bootstrap CI using 1cM blocks. For several populations in the tails of the cline, we do not show A and/or C within-ancestry estimates because these populations have too few high-confidence ancestry blocks for accurate estimation (see methods). The low diversity outlier at 33.34 degrees latitude in the N. American cline is the 2014 Avalon sample, which comes from a small island population off the coast of California.

Scutellata-European hybrid bee populations are consistently more diverse than reference bee populations because they are genetic mixtures of these diverged groups. We do, however, observe a drop in diversity in the tails of both hybrid zones starting at approximately 34.5°latitude from the equator. We tested whether this drop in diversity is necessarily the result of a bottleneck or can be explained solely by a cline in mean ancestry composition from more diverse *scutellata* and highly admixed genomes to less diverse European stock. To test this alternative, we predicted population diversity from a simple weighted average of A, C, and M reference allele frequencies and

the observed population ancestry proportions. We find that based on ancestry composition alone, we do expect a drop in diversity across the hybrid zones, although the observed drop is slightly less than our predictions (Fig 1.32).

Levels of diversity within European C and M ancestries are similar to the reference panels and stable across latitude, evidence that a diverse population of European ancestry bees hybridized with *scutellata* bees as they expanded away from Brazil. We also find high diversity within A haplotypes in both hybrid zones, again consistent with no bottleneck associated with the rapid expansion. However, the diversity in the A ancestry background does decline in populations furthest from the equator, which is consistent with either strong filtering of *scutellata* haplotypes by selection or stochastic haplotype loss due to small *scutellata*-ancestry population sizes in the tails of the clines.

Scan for ancestry-associated selection. We identified loci with unusually high A ancestry frequencies, a signal of natural selection, using our MVN simulations of background covariance in ancestry to set a false discovery rate. The ancestry covariances are important to account for when testing for putative selected loci that depart from genome-wide background ancestry patterns, because deviations in ancestry are correlated across populations. Although many population pairs have only small positive ancestry covariances, the cumulative effect on the tails of the distribution of A ancestry frequencies in the larger sample is striking. These covariances can confound outlier tests for selection which only consider variance from sampling (e.g. Poisson-binomial, e.g. [**37**]). We find that by incorporating background patterns of shared drift (or weak genome-wide selection) into our null model, we can match the bulk of the observed ancestry distributions across the genome (Fig 1.5).

Loci important to the successful invasion of *scutellata*-European hybrid honey bees are likely to have an excess or deficit of *scutellata* ancestry across both continents. Thus, we tested separately for high and low A ancestry outliers on each continent, and then identified overlapping outliers between the two hybrid zones. We find evidence of selection favoring *scutellata* ancestry at 0.34% of loci in N. America and 0.13% of loci in S. America, across 14 chromosomes (Fig 1.6A). From these outliers, we find 13 regions with an excess of A ancestry in both hybrid zones at less than a 10% false-discovery-rate (top right corner of Fig 1.5). The majority (11/13) of shared outliers co-localize within a $\sim$1.5Mb region on chromosome 1, but within this region outliers separate into

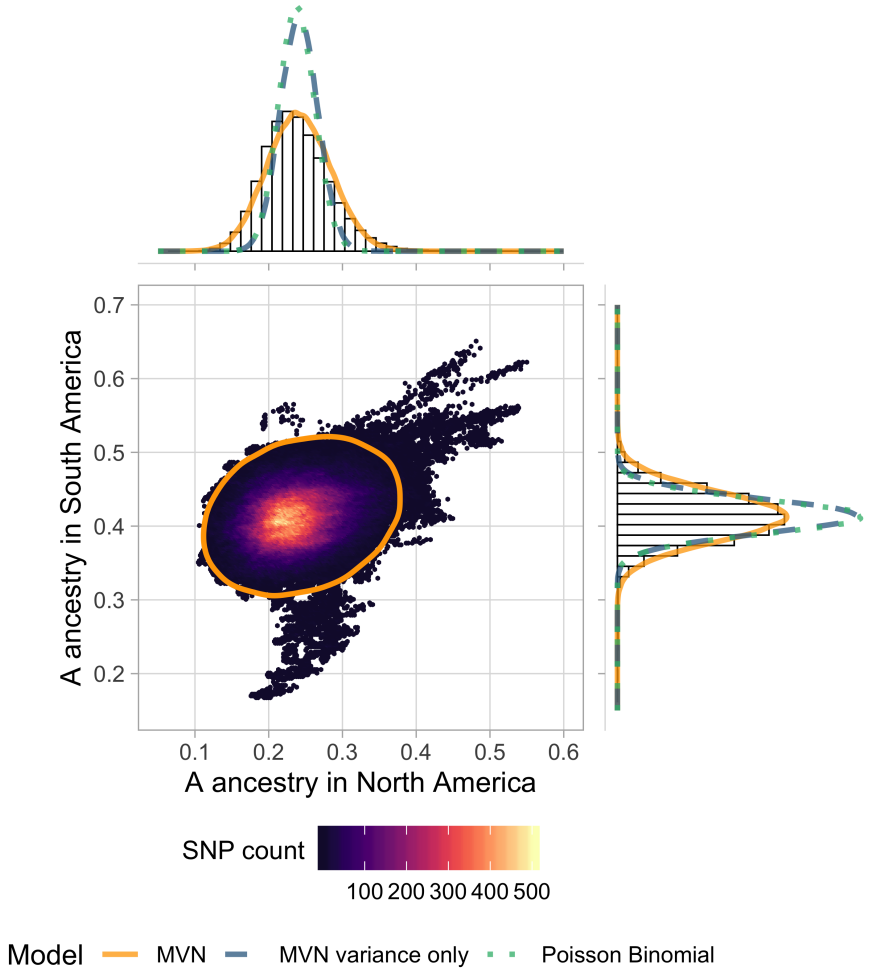


FIGURE 1.5. **Local ancestry outliers compared across hybrid zones.** We plot mean A ancestry frequency in North vs. South America for 425k SNPs across the honey bee genome. SNPs are binned for visualization, and colored by the number of SNPs within each hexagon. The orange ellipse shows the approximate 99% highest posterior density interval (HPDI) based on the full MVN model, which accounts for drift in ancestry both within and between populations. Using the same axes, we show the marginal histograms of A ancestry for each continent separately (top and right panels). Imposed on these histograms we plot density curves for 3 possible null distributions for ancestry frequencies: the full MVN model, a variance-only MVN model which only accounts for drift within populations, and a Poisson binomial model which only includes sampling variance. Most of the genome is consistent with neutrality under a MVN normal model of drift (98.6% of SNPs fall within the orange ellipse), but there are also some clear outliers. SNPs in the top right, with higher than expected A ancestry proportions in both hybrid zones, are our best candidates for loci underlying adaptive *scutellata*-ancestry associated traits. Note: While SNPs are thinned for LD, large outlier regions span many SNPs, which creates the streak-like patterns in the scatterplot.

multiple distinct peaks (Fig 1.6B). One way a cluster of A ancestry peaks could form is if favored *scutellata* alleles experience additional indirect selection from being in linkage disequilibrium with other favored *scutellata* alleles at nearby loci, thereby increasing the total effective selection in a region [**70**]. While ancestry-informative markers (AIMs) with fixed or nearly fixed differences between *scutellata* (A) and both European (C & M) ancestries are relatively rare, we were able to confirm the highest A peak within this cluster using AIMs outside of the local ancestry inference SNP set (Fig 1.27). A alleles at this main peak appear to have introgressed to high frequency hundreds of kilometers past the hybrid zone centers in both North and South America, but not reached fixation in any population (Fig 1.7). The rapid rise and slow fixation of A ancestry at this locus is potentially consistent with dominant fitness benefits. How far these A alleles have introgressed past the present hybrid zones is currently unknown because they exceed our range of sampling.

Our goal was to identify regions of the genome where high fitness is broadly associated with *scutellata* ancestry, but an alternative explanation for high A ancestry at a locus is that a very recent adaptive mutation just happened to fall on an A haplotype, initiating a classic 'hard sweep'. For shared high A-ancestry outlier regions, we distinguished between these two scenarios using population differentiation (F_{ST}) within A ancestry. We analyzed differentiation across the large cluster of shared high A-ancestry outlier peaks on chromosomes 1 and across a smaller region on chromosome 11 that contains the other two high A ancestry outliers shared between continents. We did not find high allelic differentiation between North and South American A ancestry tracks and the *scutellata* A reference panel from Africa (Figs 1.29 and 1.30), suggesting that *scutellata* ancestry in general, and not one particular haplotype, was favored by natural selection at these loci.

We used previous literature and gene orthologs to identify possible adaptive functions for regions of the genome where selection has favored *scutellata* ancestry. There are 3 major quantitative trait loci (QTLs) associated with defense behaviors (e.g. stinging) in genetic crosses of defensive *scutellata*-European hybrid honey bee colonies and low-defense European colonies [**15**, **72**], none of which overlap any signatures of selection from this study. No studies have mapped the genetic basis of elevated *Varroa* defense in *scutellata*-European hybrid (vs. European) bees, but we were

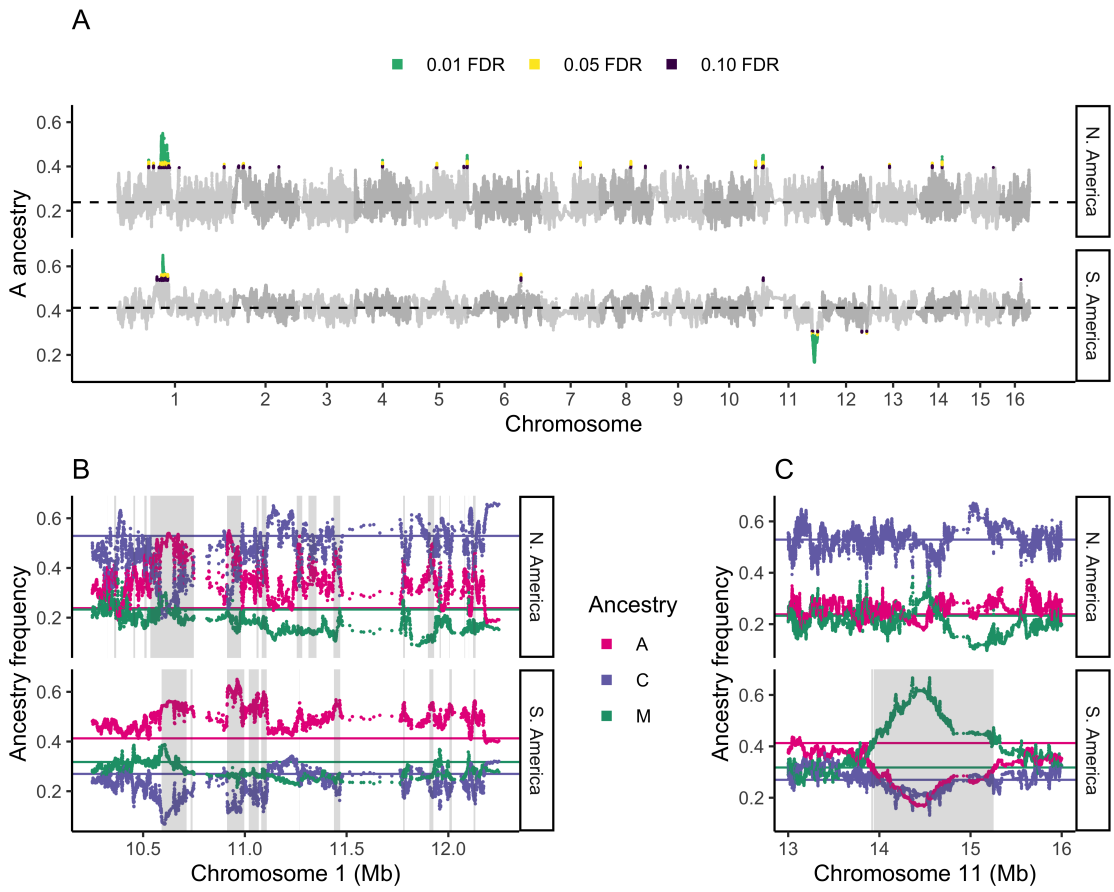


FIGURE 1.6. **Genomic location of ancestry outliers.** (A) Mean *scutellata* (A) ancestry in each hybrid zone at SNPs across the genome, with outliers colored by their false-discovery-rate. Genome-wide mean A ancestry in each zone is indicated with a dashed line. Shared peaks for high A ancestry are seen on chromosomes 1 and 11; there are no shared peaks for low A ancestry. (B) Zoomed in view of cluster of shared high A ancestry outliers on chromosome 1, with European ancestry separated into eastern (C) and western (M) subtypes. Genome-wide mean frequencies for each ancestry are shown with colored lines. Outlier regions meeting a 10% FDR for high A ancestry are shaded in grey. Shared outliers between continents overlap between the top and bottom panels. (C) Zoomed in view of the high M European ancestry outlier region found in South America. Outlier regions for low A ancestry (<10% FDR) are shaded in grey. Note: The x-axis scale differs between plots.

able to compare our results to quantitative trait loci (QTLs) associated with anti-*Varroa* hygiene behaviors [**73**, **74**, **75**] and defensive grooming [**76**] more generally. The cluster of peaks for high shared A ancestry on chromosome 1 overlaps a putative QTL associated with removal of *Varroa*-infested brood [**75**], but there are a number of large QTLs in the genome. A total of 104 genes

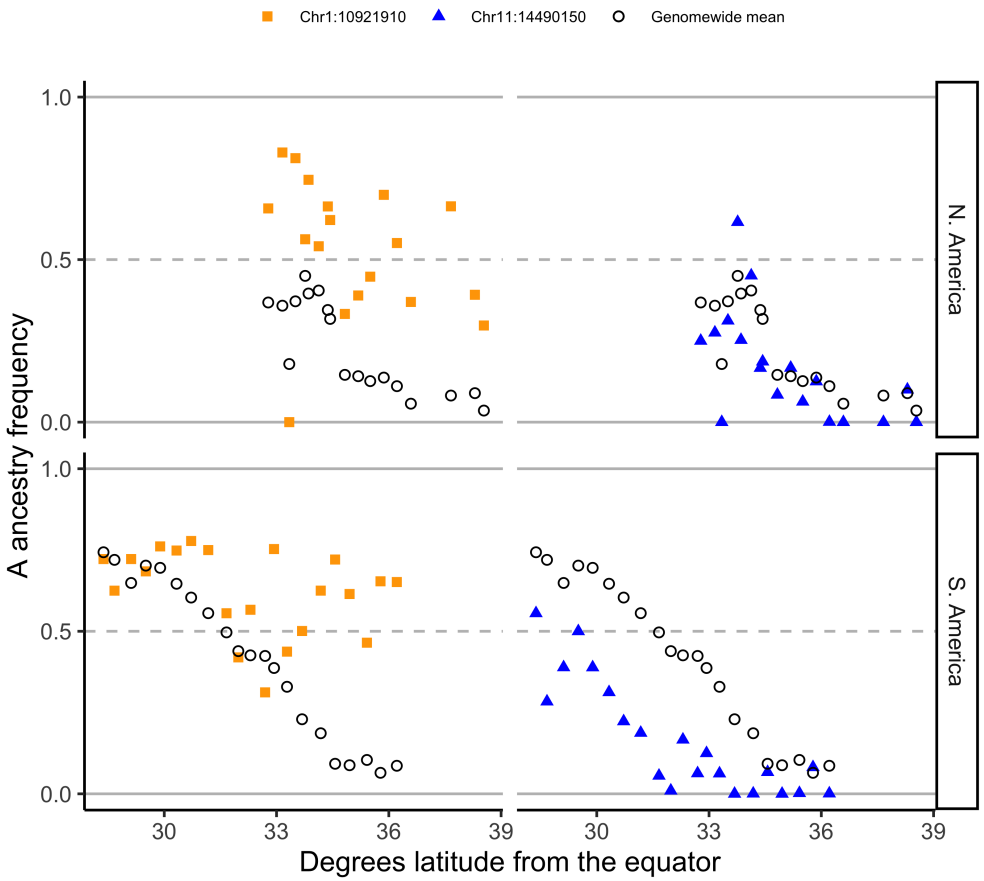


FIGURE 1.7. **Ancestry clines at outlier SNPs.** At two top outlier SNPs, we show the clines for mean population *scutellata* (A) ancestry across latitude in North (top) and South (bottom) America. To the left, we show the ancestry cline for the SNP with the highest A ancestry in the combined sample, located within the top peak on chromosome 1 for high shared A ancestry across continents. To the right, we show the ancestry cline for the SNP with the lowest A ancestry in South America, located within the large outlier region on chromosome 11 for high M and low A ancestry. Genome-wide mean local ancestry calls for each population are shown for comparison as black circles.

overlap high A ancestry outlier peaks. Predicted functions for these genes (primarily based on fly orthologs) are not significantly enriched for any Gene Ontology (GO) categories, which may simply reflect that many outlier regions are broad and contain many genes, most of which are likely unrelated to their rise in ancestry frequency. A smaller set of 37 genes with high A ancestry have signatures of selection on both continents. For these, we searched the literature and found

that two have been associated with *Varroa* in previous studies: a myoneurin (LOC725494) that is overexpressed in the brains of *Varroa* infected worker bees compared to *Nosema* infected bees and controls [**77**] and an uncharacterized protein (LOC725683) that is over-expressed in parasitized drones compared to non-parasitized drones [**78**]. While these are potentially intriguing candidates for selection, *Varroa* is only one of many possible selective pressures, and more work is needed to link the signals of selection we find here to adaptive functions.

In contrast to high A ancestry outliers, we do not find any shared outliers for European ancestry. However, the most striking example of a single-zone selection event is a large 1.4Mb region on chromosome 11 with excess European ancestry in South America (bottom middle of Fig 1.5). This region was previously identified to have low A ancestry in *scutellata*-European hybrid honey bee populations from Brazil [**37**]. We independently identify this region as a low-A ancestry outlier across Argentina, using different A/C/M reference bees, ancestry calling algorithm, and bee samples than the previous paper. We find that populations across the South American hybrid zone have reduced A ancestry at this locus, but that North American populations do not appear to have experienced selection (Fig 1.7). By including C-lineage diversity in our admixture analysis, we additionally show that this region is specifically elevated for M haplotypes, and not European haplotypes more broadly (Fig 1.6C). This region has many diverged SNPs between the three ancestry groups beyond the SNPs included in our ancestry_hmm analysis, which we use to confirm high rates of M introgression (Fig 1.28). It does not appear that a new mutation or narrow set of haplotypes was favored within M because we see little differentiation between the M ancestry in this selected region compared to the M reference panel. Additionally, within the selected region we find a peak of high F_{ST} between A, C, and M reference panels (Fig 1.31), which is consistent with this region having historically been under selection within these ancestry groups. Finally, the California bees do not have a significant deficit of A ancestry like the Argentinian bees do, but they do have two narrow peaks of excess M ancestry within this region, in the top 3% and 7% empirical percentiles for M ancestry genome-wide. Our data support a scenario in which a diversity of M haplotypes carrying the favored allele were driven to high frequency in South America after *scutellata*-European hybrid honey bees spread north of Brazil. Potential candidate genes specific to this large M-ancestry outlier region on chromosome 11 are previously described by [**37**]. In total,

we find 186 genes that overlap low-A ancestry outlier peaks (<10% FDR). These genes are not functionally enriched for any Gene Ontology (GO) categories.

Discussion

The introduction of *scutellata* honey bees to Brazil in the 1950s sparked one of the largest and best studied biological invasions known to date, with *scutellata*-European hybrids spreading from a single point of release over much of the Americas in less than 50 years. We add to this literature the first comparative study of the invasions in North and South America, with genome-wide resolution on the present distribution of *scutellata* (A) ancestry.

The parallel alignment of the genome-wide cline with latitude in both continents, despite very different length dispersal routes, strongly supports the view that *scutellata* ancestry has reached a stable climatic range limit. Because our transect in California only covered the upper half of the North American cline, the full shape for this genome-wide cline is uncertain, and may be asymmetrical because one signature of a moving hybrid zone is elongation of the lagging tail [**79**]. In contrast, we have strong evidence for convergence in the low-A portions of these two genome-wide clines, which are not expected to be distorted by cline movement, and reflects similar latitudinal range limits for *scutellata* ancestry in North and South America. Global warming trends could shift the location of the observed clines towards the poles, as has been documented for other hybrid zones sensitive to climate change [**80**]. While we currently lack temporal data with comparable genomic and geographic resolution, our results can be used as a baseline for future study.

Significant effort has been focused on finding an environmental isocline that divides regions where *scutellata*-European hybrid honey bees are expected to dominate and regions where they cannot overwinter (e.g. [**23**, **24**, **25**, **26**]). However, we observe ancestry clines that are hundreds of kilometers wide, not the narrow clines created by strong selection across a discrete environmental transition. Theoretically, these broad clines could be consistent with neutral diffusion of ancestry by migration over tens of generations. However, a scenario of neutral diffusion is inconsistent with external evidence of the rapid spread of the invasion and strong fitness trade-offs with climate: the high competitive advantage and very rapid advance of *scutellata*-associated traits and A mitochondria in the tropics in the face of considerable interbreeding with European bees and, conversely,

documented low fitness of *scutellata*-European hybrids in cooler climates, with low overwinter survival and maladaptive metabolic efficiency, foraging preferences and nesting behaviors (see [**24**] for a review). Thus, we conclude that honey bee fitness is more likely to be tracking environmental variables with smooth transitions over broad geographic regions (e.g. climate), which may create intermediate environments where ancestry intermediates have higher fitness, thus broadening the observed hybrid zones. These proposed dynamics are similar to well-studied cases in other systems where bounded hybrid superiority and/or local adaptation to continuous environments maintain adaptive clines across broad geographic regions (e.g. [**81**, **82**, **83**]).

As a null model, we expect phenotypic clines to match the scale of the observed ancestry clines, with smooth transitions in mean phenotype over hundreds of kilometers. Many phenotypes of interest, e.g. defensive behavior or *Varroa* tolerance, are expressed or measured at the colony-level and so we could not assess these in our survey of freely foraging bees. Future phenotypic surveys could be compared with our genomic clines to ascertain if key phenotypes diverge from this expected pattern, e.g. due to strong selection beyond that experienced by the rest of the genome. Indeed we see that wing length, a trait hypothesized to be associated with latitudinal body-size adaptation following Bergmann's rule [**64**], has a geographic distribution consistent with the genome-wide ancestry cline. This suggests that while wing length, which is strongly correlated with body size [**46**, **84**], may well have fitness trade-offs with climate, selection for these traits does not appear to be strong enough compared to average selection for ancestry to deviate from background genomic patterns over a short time scale.

We observe relative uniformity at the climate barrier, with no individual loci showing steeper ancestry clines than what can be produced by a null model accounting for background patterns of variation in ancestry frequencies shared across populations. Nor do we observe any loci that have below 10% frequency of A ancestry in California, despite the large distance and climatic range traveled over by this portion of the invasion. If the invasive ability of *scutellata*-European hybrid honey bees were due to a small number of loci we would expect *scutellata* ancestry to have been swamped out at many unlinked neutral loci in the genome due to interbreeding at the front of the advancing wave of expansion [**85**]. Instead, relative genomic cohesion points to a polygenic basis for the high fitness and rapid spread of *scutellata* ancestry as well as the fitness costs in cooler

climates underlying the parallel range limits observed across continents. However, we note that the distinction between so-called 'Africanized' and 'non-Africanized' honey bees is likely to further blur over time. Genetic barriers are strengthened when selection is distributed across many loci, but they are still easily permeated by adaptive alleles [**86**]. Furthermore, given high recombination rates in honey bees, we predict only loci tightly associated with climate-based fitness trade-offs will remain geographically bounded over long periods of time.

Our findings add to the genomic evidence that *scutellata*-European hybrid honey bees cannot be treated as a single genetically and phenotypically cohesive group. We show that bees have intermediate *scutellata* ancestry proportions over large geographic areas, with no evidence that *scutellata*-European hybrid honey bees share any defining *scutellata* ancestry loci (including mtDNA). Colonies within these wide hybrid zones have largely unknown colony-defense behaviors and are likely to show high variance in many traits, overlapping with variation within European bees. These bees defy 'Africanized' (vs. 'non-Africanized') labels currently used by researchers, beekeepers, and policy makers. While more precise ancestry information is becoming increasingly available, it's important to understand the limitations for trait prediction. Importantly, there is no one-to-one mapping between A ancestry and colony defense. Recent findings show that both *scutellata* and M European ancestry contribute to defensiveness segregating in *scutellata*-European hybrid populations in Brazil [**87**]. Additionally, 'gentle Africanized honey bees' in Puerto Rico show that *scutellata*-European hybrid honey bee populations can evolve low defense while maintaining *scutellata* ancestry and other associated traits [**49**,**50**]. Future research could improve upon ancestry-based trait predictions by identifying genetic markers for agriculturally undesirable and beneficial traits segregating in *scutellata*-European hybrid honey bee populations.

Scutellata-European hybrids provide a promising source of genetic variation for breeding in light of the vulnerability of European lineages to current environmental stressors and associated bee declines [**88**]. *Scutellata*-European hybrid honey bees have high competitive fitness and, we show here, maintained high genetic diversity despite their rapid expansion. In this study, we have taken a first step towards mapping the genetic basis of the high fitness of *scutellata*-European hybrid honey bees by identifying loci where selection has favored *scutellata* or European ancestry in both North and South America. We identify several loci with convergently high A ancestry on

both continents, and many more across the genome with evidence of selection favoring A ancestry in one hybrid zone. In contrast, with the exception of one striking outlier for high M ancestry in South America, we find little evidence that European ancestry or admixture per se contributed broadly to the success of *scutellata*-European hybrid honey bees. We attribute this difference in results from a previous study of *scutellata*-European hybrid honey bees in Brazil [**37**] to a more appropriate null model that accounts for shared variance in ancestry across populations. While our population genetics approach is trait-blind, our results can be compared to future functional and genetic mapping studies to look for overlap between trait-associated and positively selected loci. Applying similar methods to other systems, especially where replicated hybrid zones can be sampled, holds great promise for revealing loci important to adaptation.

Materials and Methods

Statistical results and figures were created in R [**57**] with use of the tidyverse [**58**] packages. Other scripts were run using GNU parallel [**89**].

Sampling. We sampled individual foraging honey bees across two hybrid zones, located at the transitions to temperate climates in North and South America. We sampled at least 10 bees each from 12 populations in California and 21 populations in Argentina (see maps, Fig 1.1).

For each population, we hand-netted individual foraging bees within a sampling radius of approximately 15km. Because commercial colonies are often temporarily relocated for the spring pollination season, we sampled in summer, when foraging bees are more likely to come from resident populations. We additionally included in our analyses 35 high-coverage published genomes of freely foraging bees collected from 6 populations between September 2014 and January 2015: Davis, Stebbins, Stanislaus, Avalon (Catalina Island), Placerita, and Riverside (Sky Valley and Idyllwild) [**34**]. While these sampled bees come from an unknown mixture of local feral and domesticated colonies, previous surveys from California have found that freely foraging bees tend to closely match feral sources, based on mtDNA composition [**30**]. Consistent with this view, eight of our sequenced bees from different populations in Argentina were collected close to a feral nest (< 5m), but do not appear to be ancestry outliers for their sampling locations. More specifically, we fit a general linear model (logit(A ancestry) $\sim$ absolute latitude + feral nest) using glm with

gaussian errors in R and found no significant effect on A ancestry of sampling near a feral nest ($P = 0.97$). Based on these results and our seasonal timing, the bees in this study are likely sourced primarily from local feral populations, with some contribution from resident domesticated bee colonies.

Lab work and sequencing. We selected a subset of 279 bees from our North and South American hybrid zones for whole genome sequencing, 8-9 bees per sampled population (see Table 1.1). For each bee, we dissected wing flight muscles from the thorax and extracted DNA using QIAGEN DNeasy Blood and Tissue kits. We followed a new high-throughput low-volume DNA library preparation protocol (see [**90**] for details, "Nextera Low Input, Transposase Enabled protocol"). Briefly, we prepared individual Nextera whole-genome shotgun-sequencing DNA libraries using enzymatic sheering and tagmentation. Then we PCR-amplified and barcoded individual libraries using the Kapa2G Robust PCR kit and unique custom 9bp 3' indices. Finally, we pooled libraries within each lane and ran bead-based size-selection for 300-500bp target insert sizes. We targeted 4-6x coverage per bee based on a preliminary analysis of our power to replicate local ancestry calls from one of the published high coverage populations (Riverside 2014) using simulated low coverage data (Fig 1.11). We multiplexed our samples across 5 Illumina HiSeq4000 lanes for paired-end 2 x 150bp sequencing. In total, we generated 5.1x mean coverage per bee for 278 samples. The 279th sample was excluded from all analyses for having extremely low (<0.1x) sequence coverage.

Alignment and SNP set. In addition to the sequence data produced by this study, we downloaded Illumina raw read sequences for 35 previously published California genomes (PRJNA385500 [**34**]) and a high-quality reference panel of *A. m. scutellata* (A, n = 17), *A. m. carnica* (C, n = 9), and *A. m. mellifera* and *A. m. iberiensis* (M, n = 9) honey bee genomes (PRJNA216922 [**91**] and PRJNA294105 [**8**]) from the NCBI Short Read Archive. For all bees, we mapped raw reads to the honey bee reference genome HAv3.1 [**92**] using Bowtie2 very-sensitive-alignment with default parameters [**93**]. We then marked and removed duplicate reads with PICARD and capped base quality scores using the 'extended BAQ' option in SAMtools [**94**]. Using the software ANGSD [**95**], we identified a set of SNPs with minor allele frequency $\geq 5\%$ in the combined sample based on read counts (-doMajorMinor 2 -doCounts 1). We excluded unplaced

scaffolds (<5Mb total) and applied standard quality filters for SNP calling (base quality $\geq$ 20, mapping quality $\geq$ 30, total read depth $\leq$ 5500 ($\sim$2x mean), and coverage across individuals $\geq$ 50%). We calculated the genetic position (cM) for each SNP using a 10kb-scale recombination map [**96**] and linear interpolation in R (approxfun). We assumed constant recombination rates within windows and extrapolated positions beyond the map using the recombination rate from the nearest mapped window on that chromosome.

We identified SNPs on the mitochondria (HAv3.1 scaffold NC_001566.1) using the same pipeline as nuclear DNA above, but allowing for extra read depth (up to 100000000x). We then called consensus haploid genotypes at these SNPs for all individuals using ANGSD (-dohaplocall 2 -remove_bads 1 -minMapQ 30 -minQ 20 -doCounts 1 -minMinor 2 -maxMis 174).

Global ancestry inference. We estimated genome-wide ancestry proportions for each bee using methods designed for low-coverage sequence data. Briefly, we combined bee genomes from the hybrid zones with reference genomes for *scutellata* (A), eastern European (C) and western European (M) bees. To reduce linkage disequilibrium (non-independence) between our markers for global ancestry inference, we thinned to every 250th SNP ($\sim$14k SNPs at 19kb mean spacing) before calculating genotype-likelihoods for each bee using the SAMtools method in ANGSD (-GL 1). We first ran a principal components analysis in PCAngsd [**97**] to confirm that genetic diversity in the hybrid zones is well-described by 3-way admixture between A, C, and M reference panels (Fig 1.8). We then estimated genome-wide ancestry proportions for all bees using NGSAdmix (K = 3) [**52**].

Local ancestry inference. We inferred the mosaic of *scutellata* vs. European ancestry across the genome of each bee using a hidden Markov inference method that can account for *scutellata* (A), eastern European (C) and western European (M) sources of ancestry within low-coverage *scutellata*-European hybrid honey bee genomes (ancestry_hmm v0.94 [**53**]). For local ancestry inference, we enriched for ancestry-informative sites by filtering for $\geq$ 0.3 frequency in one or more reference population (A, C, or M) and at least 6 individuals with data from each reference population. We subsequently thinned markers to 0.005cM spacing, because at that distance linkage disequilibrium within ancestries is expected to be low ($r^2 < 0.2$ [**33**]), leaving a final set of 542,655

sites for ancestry calling, or $\sim 1/7$ of the original SNP set. Individual bees sequenced in this study and previously published California bees have 5.42x and 14.5x mean coverage, respectively, across this final SNP set. For each population, we jointly estimated time since admixture and ancestry across the genome of each individual, using read counts from the hybrid zone and allele frequencies for A, C and M reference populations at each SNP. To generate major/minor allele counts for each reference population, we used ANGSD to call genotypes (-doPost 1) using a minor allele frequency prior (-doMaf 1) and the SAMtools genotype likelihood (-GL 1), after quality filtering (map quality ≥ 30, reads matching major/minor allele $\geq 60\%$, and read depth ≥ 6x). As additional inputs to ancestry_hmm, we used NGSAdmix results as a prior for population ancestry proportions and set the effective population size to $N_e = 670,000$ [**37**]. We modelled a simple three-way admixture scenario: starting with C ancestry, we allowed for a migration pulse from M and a second, more recent, migration pulse from A. Timing of both migration pulses were inferred from the range 2-150 generations, with priors set at 100 and 60 generations. To calculate a point estimate for each individual's ancestry proportion at a SNP, we marginalized over the posterior probabilities for homozygous and heterozygous ancestry from the ancestry_hmm output (i.e. $A = p(AA) + 1/2(p(CA) + p(MA))$.

Ancestry covariance matrix. To explore how populations vary and covary in their *scutellata* ancestry along the genome we calculated the empirical population ancestry variance-covariance matrix (K), an admixture analog of a genotype coancestry matrix (e.g. [**98**]). The K matrix is calculated using population *scutellata* (A) ancestry frequencies inferred by the local ancestry HMM, e.g. for populations i and j with mean ancestry proportions α_i and α_j, and ancestry frequencies at a locus $anc_{i,l}$ and $anc_{j,l}$, their ancestry covariance calculated across all L loci genome-wide is

$$K[i,j] = \frac{1}{L}\sum_{l=1}^{L}(anc_{i,l} - \alpha_i)(anc_{j,l} - \alpha_j).$$

Ancestry correlations between the high-latitude cline endpoints. To more formally test for excess ancestry correlations between more geographically distant (but climatically similar) populations, we grouped populations by dividing each hybrid zone into low- and high-A ancestry regions relative to the estimated latitude for the genome-wide cline center. The southernmost 11

(out of 20) of the South American populations, and all of the sampled North American populations, fall in the 'low A' half of their respective hybrid zones. We calculated mean ancestry covariances (K matrices) separately for each chromosome, using the genome-wide mean ancestry as α, then summarised across populations by taking the mean correlation for each type of pairwise comparison, within and between continents and regions. We tested whether, on average across chromosomes, low-A South American populations share higher ancestry correlations with low-A North American populations than with geographically closer high-A South American populations and repeated this test excluding chromosomes 1 and 11 which contain large outlier regions (Fig 1.21). We also tested the same group comparison across recombination rate quintiles instead of chromosomes (Fig 1.22).

To investigate whether recent long-distance migration likely generated the elevated ancestry correlations we observe between low-A South America and low-A North America, we looked at patterns of allelic covariance within ancestry. Specifically, for each ancestry we estimated a genetic covariance matrix in PCAngsd for all individuals sampled from the hybrid zone, based on allelic diversity within high-confidence homozygous ancestry tracts (posterior >0.8). Under recent migration, we would expect the excess A ancestry correlations between the two ends of the hybrid zones to be mirrored by allelic covariances within all three ancestries. Instead, we find that the two most prevalent ancestries, A and C, both have low or negative genetic covariances between continents (Fig 1.24). In contrast, M ancestry does show an excess of cross-continent covariance, and we followed up to determine if this is uniquely American covariance (i.e. the result of shared drift within the Americas) or could have been imported from Europe. Adding reference individuals to these within-ancestry analyses, we find that M ancestry in the Americas imported pre-existing structure from Europe, with more Poland-like (*Apis mellifera mellifera*) than Spain-like (*Apis mellifera iberiensis*) M ancestry at the temperate ends of the clines (Figs 1.25 and 1.26).

Simulated ancestry frequencies. At various points in the results we compare our outliers to those generated by genome-wide null models of ancestry variation along the genome. We simulated variation in ancestry frequencies at SNPs across the genome under three models: (1) A Poisson-binomial model that only accounts for sampling variance, not drift (e.g. [**37**]); (2) a multivariate-normal model with covariances set to zero, which accounts for effects of both sampling and drift within-populations (e.g. [**99**]); and (3) a multivariate-normal model with covariances to additionally

account for shared drift in ancestry between populations. For each model, we simulated in R neutral A ancestry frequencies at 100,000 independent loci [**100**, **101**]. The full multivariate-normal model is used for comparison to the results, while the first two models are only used to show the effects of ignoring covariances.

In the Poisson-binomial simulation, for each bee we sampled 2 alleles from a binomial distribution with mean equal to the individual's genome-wide ancestry proportion inferred by ancestry_hmm.

For the variance-only MVN simulation, we empirically calibrated an independent normal distribution for each population that can exceed binomial ancestry variance (e.g. due to drift). This model is equivalent to the full MVN model below, but sets all off-diagonal entries of the K ancestry variance-covariance matrix to zero.

In our full multivariate-normal model, we account for non-independent ancestry within and between our sampled populations:

$$\text{A ancestry} \sim \text{MVN}(\alpha, K),$$

where α is the vector of genome-wide mean population ancestry frequencies and K is the empirical population ancestry variance-covariance matrix. Because the MVN models are not bounded by 0 and 1, but real frequency data is, we set all simulated individual population frequencies exceeding those bounds (5.2% low and 0.09% high) to the bound. Truncation has little effect on the distribution in general and no effect on the frequency of high A ancestry outliers, but does make extremely low outliers (attributed to some populations having simulated negative frequencies) less likely (Fig 1.18). For more details on model approximations to the observed data, see Figs 1.19 and 1.20.

Cline models. To better understand the role of dispersal and selection maintaining the current geographic range limits of *scutellata* ancestry, we fit a logistic cline model to the individual genome-wide ancestry proportions estimated by NGSAdmix. We estimated continent-specific c and b parameters to test for a difference in cline center (degrees latitude from the equator) and/or slope between the northern and southern invasions. Then we fit a joint model with a single cline to see how well absolute latitude or climate can consistently predict A ancestry frequencies across

both continents. Specifically, we tested four environmental variables that likely contribute to varying fitness across space: mean annual temperature (°C), mean temperature of the coldest quarter (°C), minimum temperature of the coldest month (°C) and mean annual precipitation (cm). We downloaded mean climate observations for 1960-1990 [**68**] from WorldClim.org at 30 second map resolution ($\approx$ 1 km^2 at the equator) and then averaged within a 5km radius around each bee's sample coordinates. We compared climate and latitude-based selection models to a neutral dispersal model, where genome-wide A ancestry is predicted solely based on the distance (km) traveled from Río Claro, São Paulo, Brazil, the point of origin for the *scutellata*-European hybrid honey bee invasion (estimated from GPS coordinates using "distGeo" in R [**102**]). For each model, we substituted latitude, distance, or climate for x_i in Eq 1.1 and we used AIC to compare model fits.

We then fit individual-SNP clines to the mean population ancestry frequencies in South America, where our samples span the full cline. We tested for individual outlier loci that may underlie a climate barrier by fitting the same logistic cline model to a set of simulated population ancestry frequencies for S. America (see MVN simulation), and comparing observed cline slopes to this null distribution. In addition, we tested for enrichment of the empirical top 5% of steep clines in regions of the genome with low recombination rates. We divided the genome into 5 equal-sized recombination rate bins ([0, 2.92], (2.92, 21.6], (21.6, 31.7], (31.7, 38.6] and (38.6, 66.9] cM/Mb) and used 10,000 block bootstraps [**103**] to calculate basic bootstrap confidence intervals for each recombination rate quintile while accounting for spatial correlation in both cline slopes and recombination rates across the genome. For the bootstrap, we divided the genome into 0.2cM blocks, we re-sampled these blocks with replacement, and for each recombination bin we calculated mean b and the proportion of SNPs in the top 5% steepest slopes from our bootstrap sample. When fitting non-linear least squares in R for both genomewide and individual snp clines, we used multiple random starting values to make sure we searched across all local minima and found the global optimum solution (nls.multstart [**104**]). Starting values were drawn from uniform distributions: b $\sim$ Unif(-5, 5) and $\sim$ Unif(min, max) across the observed range for latitude and climate variables.

Wing morphology. We imaged a slide-mounted fore wing and measured wing length to the end of the marginal cell using imageJ (Fig 1.15). We included 269 bees in the wing analysis (only

bees sequenced by this study had wings preserved and n = 9 bees were excluded for wing tatter or damage).

We also measured fore wing lengths for A, C, and M reference bees in the Oberursel Collection sampled from their native range (n = 52 [**105**]). While the effect of ancestry on wing length is similar in magnitude and direction in both datasets, we found that the mean wing lengths for the European reference bees (C & M) fell below the mean for our American bees with close to 100% European ancestry, perhaps reflecting phenotypic plasticity. Thus we do not incorporate these measurements of A, C, and M reference bees into the subsequent analyses.

We tested various models of the relationship between wing length, ancestry and geography. First, we fit a linear model to predict wing length in our sample from genome-wide ancestry. We visually compared our wing measurements to what we would expect if the cline in wing length across latitude were fully described by this linear relationship between ancestry and wing length and our best-fit clines for genome-wide ancestry (Fig 1.2). We additionally tested for differences between continents by adding a main effect and an interaction term for South America to our linear model.

We performed admixture mapping of wing length to test if the ancestry state at any individual SNP predicts residual variation in wing length. To do this, we first regressed wing length on genome-wide A ancestry, to correct for background ancestry effects. We then took the residual wing lengths from this linear model fit and regressed these on A ancestry allele counts at each locus in turn (using the maximum a posterior probability (MAP) estimates from the local ancestry HMM). We set a genome-wide significance threshold of $p < 1.1 \times 10^{-6}$ to control for multiple testing at a 5% family-wise error rate, using an analytical approximation for admixture mapping, calculated assuming 47.6 generations since admixture [**106**, **107**].

Identifying ancestry outlier regions and genes. To identify loci underlying ancestry-associated fitness differences, we tested SNPs for an excess or deficit A ancestry within each hybrid zone. We calculated 1%, 5% and 10% false-discovery rates (FDR) by using our MVN simulation results to set the number of false-positives we expect under a neutral model for high and low A ancestry within each continent separately (one-tailed outlier tests). We then compared the overlap in outliers between hybrid zones to identify SNPs with signatures of selection on both continents.

In addition to local ancestry, we used ancestry-informative markers with fixed or nearly fixed differences to verify high-introgression regions. We defined ancestry informative markers as SNPs with coverage for at least 5 individuals from each reference panel and >0.95 allele frequency differences between the focal ancestry and both other ancestries. We estimated allele frequencies at each ancestry-informative marker using ANGSD, polarized SNPs so the focal ancestry has the highest MAF, and only included markers with coverage in all sampled populations. Ancestry-informative markers for A (n = 4,302) are relatively rare compared to markers for C (n = 17,384) and M (n = 15,626) because European populations each experienced a historical bottleneck differentiating them from the other two groups. Because of LD-thinning before local ancestry inference, 88% of these ancestry-informative markers were not included in the ancestry_hmm SNP set, and therefore provide separate support for high-introgression regions.

We identified a set of candidate genes that overlap regions of the genome with exceptionally high or low A ancestry (<10% FDR) using BEDtools [**108**]. For this analysis, we downloaded gene annotations for the HAv3.1 genome assembly from NCBI (accessed 7/22/19). 72 out of 104 genes overlapping high A ancestry peaks and 131 out of 186 genes overlapping low A ancestry peaks have associated BEEBASE gene IDs. For these high and low gene sets, we tested for enrichment of Gene Ontology (GO) terms compared to a background of all honey bee genes, using DAVID 6.8 [**109**] and a Benjamini-Hochberg corrected FDR of 5% [**110**]. To find out what is previously known about the 37 genes that overlap regions with high A ancestry on both continents, we conducted a literature search using the NCBI gene search tool and google.

We additionally checked if our candidate selected loci overlap regions of the genome previously associated with defensive or anti-*Varroa* behavioral traits (QTLs and associated marker sequences from [**15**,**72**,**73**,**74**,**75**,**76**,**111**,**112**,**113**]). We estimated genome coordinates for QTLs by blasting marker sequences to HAv3.1 and keeping the best BLASTn [**114**] hit with an E-value <0.01 (see Table 1.5). When assessing physical overlap between genome annotations and ancestry outliers, we assumed ancestry calls for a SNP apply to the short genomic window around that SNP, spanning midway to the next ancestry call. When visualizing and counting the number of selected regions in the genome, we further merged near-adjacent (<10kb) significant ancestry windows into contiguous regions.

Population diversity. We calculated allelic diversity (π) for each population and our A, C, and M reference panels. First, we calculated a simple unbiased population allele frequency in ANGSD based on a weighted average of observed read counts (counts -8) for each SNP. For this analysis, we included all SNPs ascertained in the combined sample (see 'Alignment and SNP set' above) but excluded SNPs from a population's estimate when fewer than two individuals had coverage. Using these allele frequency estimates, and a finite-sample size correction ($n = 2\times$ number of individuals with data at a site), we calculated mean per-SNP heterozygosity. To approximate uncertainty in our estimates, we divided the genome into 5,254 non-overlapping 1cM blocks, re-calculated our diversity estimates for 10,000 block bootstrap samples, and calculated a 95% simple bootstrap confidence interval. Finally, to get per-bp diversity, we scaled our per-SNP diversity estimates by the density of SNPs in the genome, using the same coverage and depth quality filters in ANGSD as in our SNP pipeline to count total mappable sites.

For within-ancestry diversity estimates, we used our ancestry calls to identify contiguous tracts with high posterior probabilities (>0.8) of homozygous A, C, or M ancestry. We used these tracts to divide the genome into high confidence A, C, and M ancestry states, and filter for reads that mapped within these states. We then repeated the estimation and block bootstrap procedure above using only the reads associated with a particular ancestry. To estimate within-ancestry diversity for a population, we required data for at least 75 1cM blocks spread across at least 15 of the 16 chromosomes, which excludes 6 populations with too little A ancestry in the tails of both clines and 8 populations with too little C ancestry in the S. American cline for accurate estimation. We compared observed and predicted heterozygosity for each population based on expected allele frequencies calculated by multiplying population-specific admixture proportions by reference population allele frequencies for each ancestry.

To test whether selection had favored specific haplotypes, or *scutellata* ancestry more generally, within shared outlier regions for high A ancestry, we calculated population differentiation (F_{ST}) between the A reference panel and A haplotypes within each hybrid zone. We also calculated within-ancestry F_{ST} between the two hybrid zones, to assess whether the same A haplotypes rose in frequency on both continents. Likewise, for the large high M outlier on chromosome 11, we calculated pairwise differentiation within M ancestry between North America, South America, and

the M reference panel. We similarly calculated F_{ST} for all three contrasts between A, C, and M, reference panels across these outlier regions, to test for signatures of historical selection and divergence between these ancestry groups. For F_{ST} calculations, we estimated within-ancestry allele frequencies for North and South America using the same method described above for within-ancestry π, except pooling individuals by hybrid zone rather than population. We used Hudson's estimator for F_{ST} (Eq 10 in [**115**]), calculated the average per-SNP F_{ST} within sliding 50kb windows stepping every 1kb across ancestry outlier regions, and only included SNPs with coverage for at least two individuals for both populations in the contrast and windows with at least 10 SNPs.

Ethics Statement

Honey bee samples from California were collected with permission from the California Fish and Wildlife (wildlife.ca.gov; permit ID D-0023599526-1). Honey bee samples from Argentina were collected and transferred to the University of California, Davis, for genomic analysis with authorization from Argentina's National Institute of Agricultural Technology (Instituto Nacional de Tecnología Agropecuaria, INTA Argentina, www.argentina.gob.ar/inta, document ID 25401; MTA No. 2018-0374-M filed at UC Davis). This study did not involve any endangered species, protected species or protected areas.

Acknowledgements

We thank the National Institute of Agricultural Technology (INTA), Argentina, for the use of bee samples from Argentina. Thank you to Stefan Fuchs and the Oberursel Collection for sharing wing images of A, C and M reference bees. Thank you to Brock Harpur for making marker sequences from previous Hunt lab QTL studies publicly available on Data Dryad. Thank you to Julie Cridland and Nicholas Saleh for early tips on sampling and California bees, to Philipp Brand, Brenda Cameron, and Anne Lorant for sharing protocols and advice on lab work, to Jodie Jacobs and Petra Silverman for help phenotyping wing traits, and to Kelsey Lyberger, Roisin McMullen and Jodie Jacobs for their help in the field netting bees. Thank you to Jennifer Van Wyk, Hannah Whitehead, and Ang Roell for useful conversations about how scientists and the public describe *scutellata*-European hybrid honey bees and the broader impacts. Thank you to Daniela Zarate,

Jeffrey Ross-Ibarra, Michael Turelli, Peter Ralph and to members of the Coop, Ramírez, and Ralph labs for insightful discussions and feedback on earlier drafts of the manuscript.

This work was funded by the National Institute of General Medical Sciences of the National Institutes of Health, www.nigms.nih.gov (NIH R01 GM108779 and R35 GM136290, awarded to GC), the Division of Integrative Organismal Systems from the National Science Foundation, www.nsf.gov (NSF No. 1546719, awarded to GC), the North American Pollinator Protection Campaign and Pollinator Partnership, www.pollinator.org/nappc (NAPPC Honey Bee Health Grant, awarded to EC and SR), and the Center for Population Biology UC Davis, https://cpb.ucdavis.edu (Pengelley Award, awarded to EC). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Associated Publication

Chapter 1 was published October 19, 2020:

Calfee E, Agra MN, Palacio, MA, Ramírez SR, Coop G. Selection and hybridization shaped the rapid spread of African honey bee ancestry in the Americas. PLOS Genetics. 2020; doi:10.1371/journal.pgen.1009038.

Supporting Information

Data Availability. Raw Illumina sequence data generated by this study is available through the NCBI Short Read Archive, PRJNA622776. Access information for all previously published genomic resources used in this study: HAv3.1 reference genome and gene annotations (NCBI PRJNA471592), recombination map (available by request from Jones et al.), bee genomes from California (NCBI PRJNA385500) and A, C, and M reference populations (NCBI PRJNA216922 and PRJNA294105). Bee metadata, including GPS locations and measured wing lengths are included in Table 1.1 below. Wing images generated by this study are available through Data Dryad: https://doi.org/10.25338/B8T032. Wing images for museum samples of A, C, and M bees are available by request from the Morphometric Bee Data Bank, Institut für Bienenkunde, Oberursel, Germany (https://de.institut-fuer-bienenkunde.de). All climate data was downloaded from WorldClim.org. Scripts are available at https://github.com/ecalfee/bees.

Table 1.1. **Sample information.** Geographic sampling information (population, location, date, whether collected by a feral nest), approximate sequencing coverage, global ancestry estimates and wing lengths for bees sequenced in this study and reference bees. *See supplemental file Table_1.1_Sample_information.txt*

Table 1.2. **Cline model comparison.** Model rankings between logistic cline fits for genome-wide *scutellata* (A) ancestry predicted by climate and distance variables.

	predictor	df.residual	deviance	dAIC	weight
1	Latitude	311	2.69	0.00	1.00
2	Mean temperature	311	3.63	93.60	0.00
3	Mean temperature of coldest quarter	311	5.25	208.60	0.00
4	Minimum temperature of coldest month	311	7.04	300.70	0.00
5	Distance to Sao Paulo	311	9.55	396.10	0.00
6	Annual precipitation	311	15.21	541.80	0.00

Table 1.3. **Ancestry outlier regions.** Genome coordinates for outlier regions with high or low *scutellata* (A) ancestry. Adjacent and near adjacent (within 10kb) ancestry windows meeting <10% FDR have been combined into contiguous regions and are labelled with the lowest FDR within the region. Note that shared high A outlier regions, by definition, will overlap high A South American and high A North American outlier regions, with bp and percent overlap listed. NA signifies not significant for that hybrid zone. *See supplemental file Table_1.3_Ancestry_outlier_regions.txt*

Table 1.4. **Ancestry outlier genes.** List of genes overlapping ancestry outliers at 1%, 5%, and 10% FDR thresholds. Minimum FDR for each continent listed separately. NA signifies not significant for that hybrid zone. *See supplemental file Table_1.4_Ancestry_outlier_genes.txt*

Table 1.5. **Approximate QTL coordinates HAv3.1.** Approximate coordinates (HAv3.1) for regions of the genome previously associated with defensive behaviors or *Varroa* tolerance. *See supplemental file Table_1.5_Approximate_QTL_coordinates_HAv3.1.txt*

Table 1.6. **Invasion dates and locations.** Approximate locations and dates of first arrival for the spread of *scutellata*-European hybrid honey bees as plotted in Fig 1.1. We estimated GPS coordinates for each historical observation using google maps and the available location description. *See supplemental file Table_1.6_Invastion_dates_and_locations.txt*

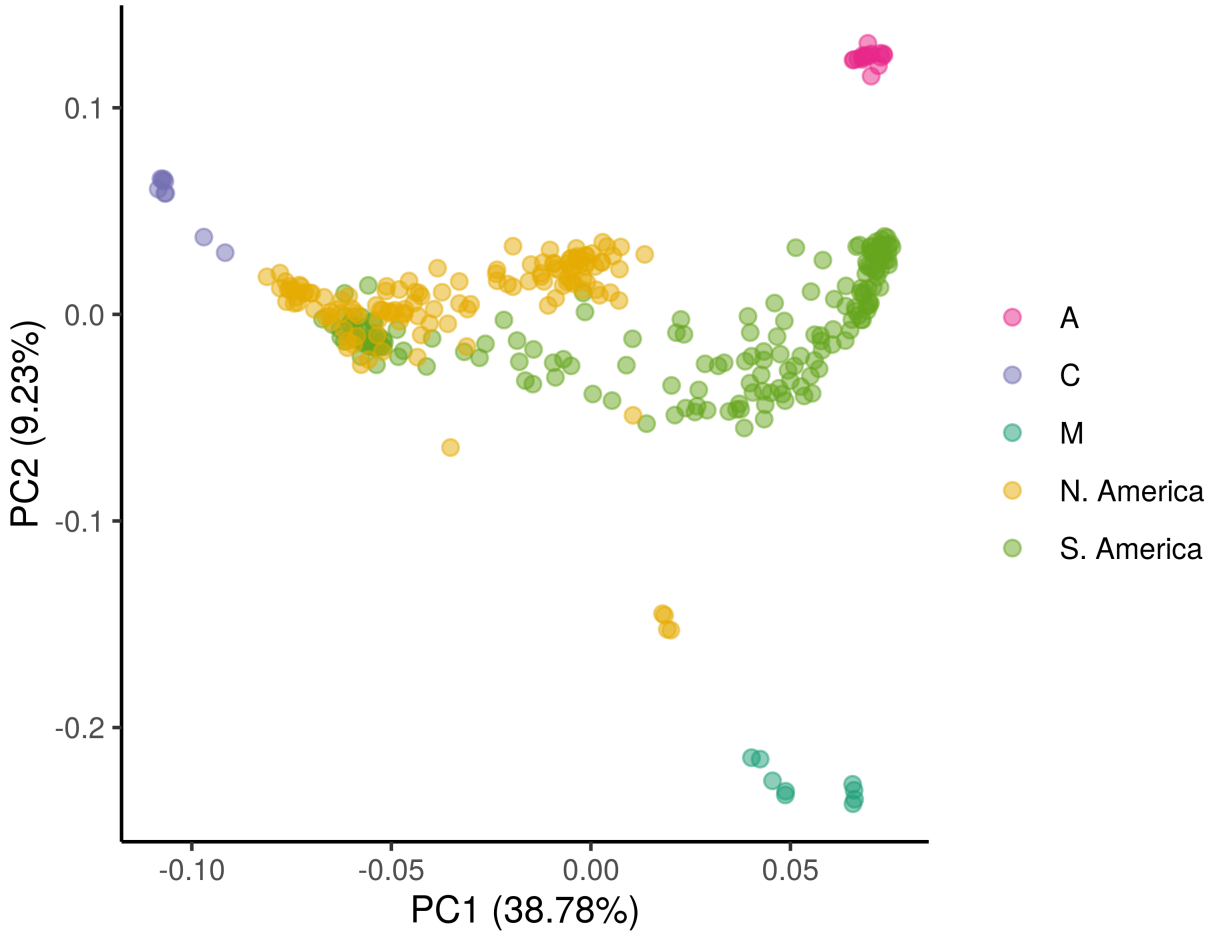


FIGURE 1.8. **PCA.** Principal components analysis generated in PCAngsd using genotype likelihoods from the same thinned set of 14,044 autosomal SNPs used in global admixture analysis. The major axes of diversity separate out C ancestry (PC1) and M ancestry (PC2). Consistent with 3-way admixture, all sampled bees from North and South America are intermediate on the PCA, in the triangle formed by reference panels for *Apis mellifera scutellata* from southern and eastern Africa (A), *A. m. carnica* from eastern Europe (C) and *A. m. mellifera* and *A. m. iberiensis* from western Europe (M).

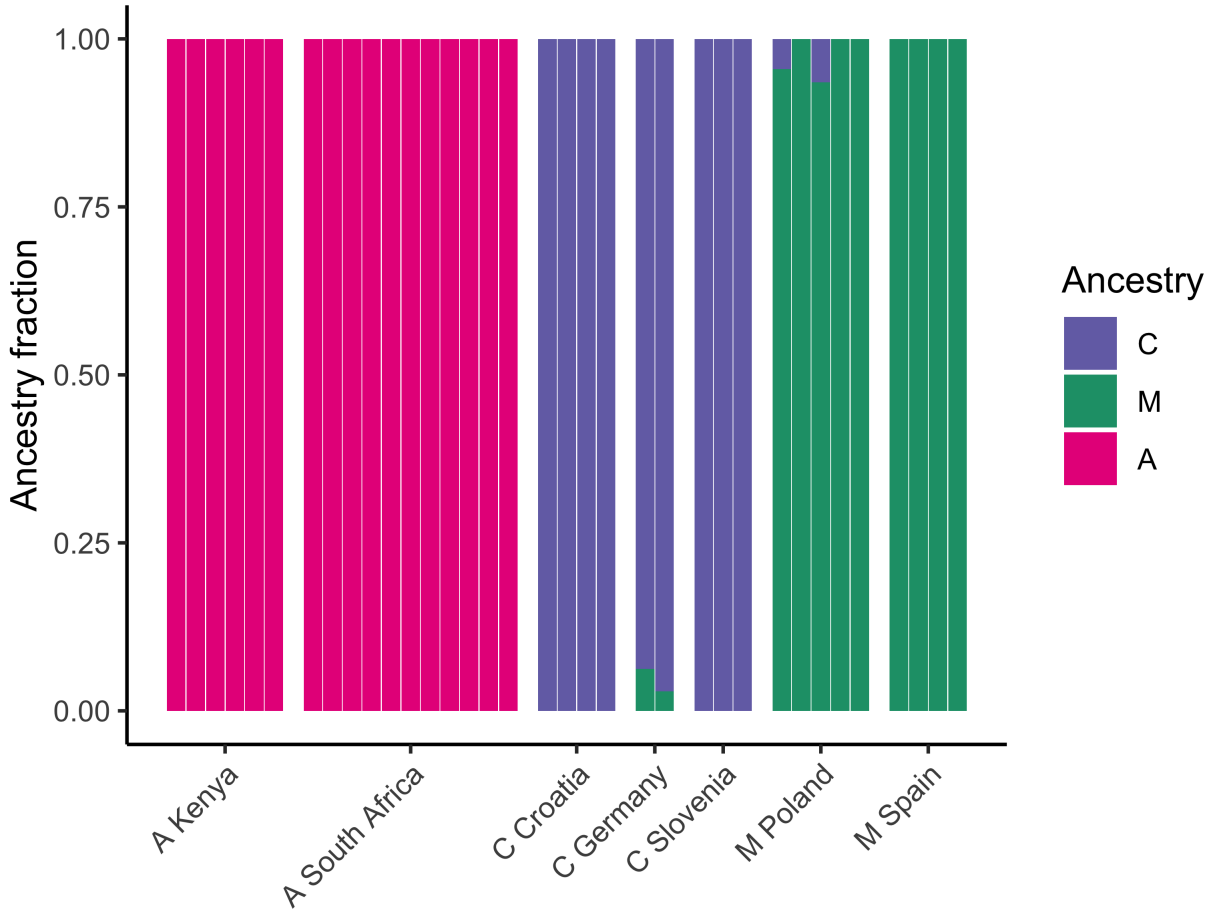


FIGURE 1.9. **Ancestry in reference panels.** Results of NGSAdmix global admixture analysis for reference populations from the combined analysis of all populations (K = 3). These results were used to assign the unlabelled ancestry components output by NGSAdmix to A, C, and M groups, based on a clear mapping to the three reference populations. We see a small amount of admixture between C and M within our reference populations, which is consistent with limited gene flow from secondary contact within Europe.

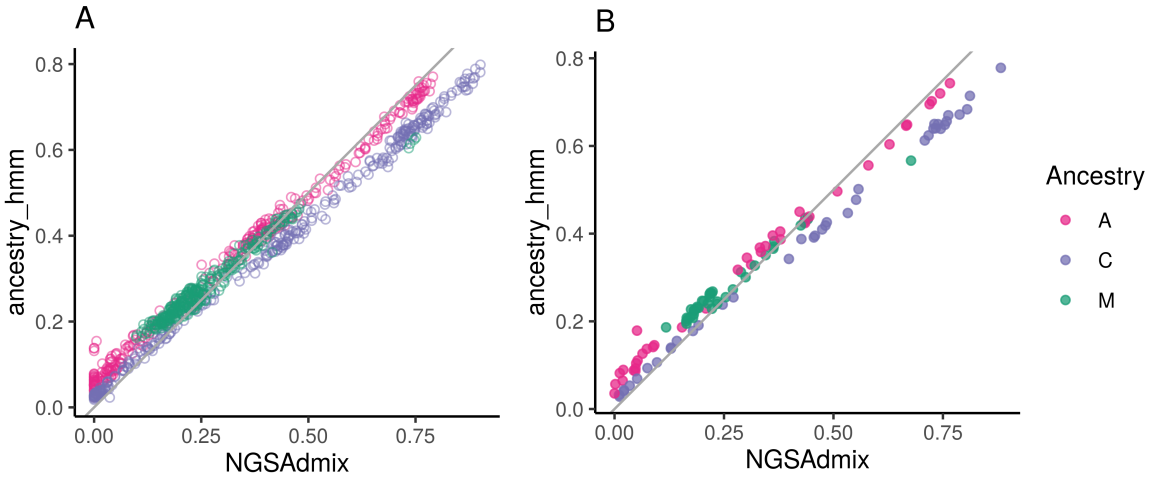


FIGURE 1.10. **Comparison of local and global ancestry results.** (A) Comparison of the mean genome-wide ancestry estimate from NGSAdmix (x-axis) and ancestry_hmm (y-axis) for each bee, with one-to-one line drawn in grey. The mean for the HMM is calculated by marginalizing the posterior over all ancestry states and taking a mean across SNPs. The individual-level ancestry estimates between the two methods agree strongly (Pearson's correlation: 0.997 A, 0.999 C, 0.985 M), but the HMM estimates slightly higher minor ancestry for bees with low admixture proportions. (B) Population mean summarises for the same comparison of NGSAdmix vs. ancestry_hmm genome-wide ancestry estimates, with one-to-one line drawn in gray. Because the population mean ancestry proportions from NGSAdmix are used as a prior for the population-specific mixing proportions in ancestry_hmm, this panel can also be interpreted as the prior (x-axis) and posterior (y-axis) of the local ancestry HMM for population-level admixture proportions.

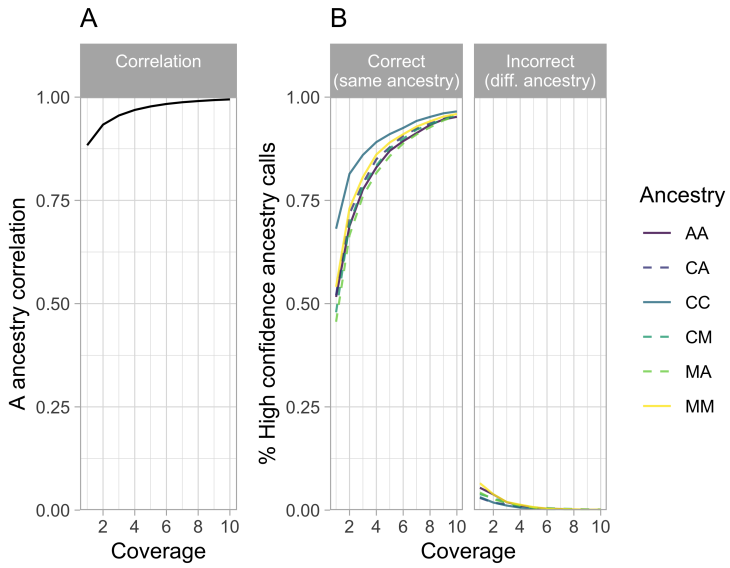


FIGURE 1.11. **Power to call local ancestry.** (A) Correlation between high-coverage and low-coverage ancestry calls, across different simulated depths of coverage (1-10x). (B) Proportion of high-confidence ancestry calls from high-coverage data that were replicated in analyses of low-coverage data, with different simulated depths of coverage (1-10x). These results are from a preliminary analysis of the power to call local ancestry accurately, used to inform target sequencing depth for this study. For this preliminary study, we used a published SNP set with data for A, C, and M reference populations [**8**] based on earlier versions of the honey bee genome (Amel4.5 [**116**]) and recombination map [**33**]. We enriched for ancestry-informativeness and thinned for linkage disequilibrium ($\geq$ 0.2 MAF in at least one reference population and $r^2 < 0.4$ within the A reference population), leaving 161k SNPs. First we ran ancestry_hmm [**53**] using called genotypes from a high-coverage admixed population with intermediate admixture proportions (Riverside 2014 (n=8): 40% C, 20% M, 40% A ancestry). We simulated lower coverage data from this same population by generating a binomial sample of n reads for each locus, based on the individual's genotype. To simulate realistic variance in coverage across the genome, n for each site and individual was generated from a negative binomial distribution with variance 3x the mean [**117**]. We additionally simulated a 1% sequencing error rate. Running local ancestry inference on the high coverage data, we inferred high confidence ancestry states for 81% of sites. First we calculated a point estimates for A ancestry ($p(AA) + 1/2(p(CA) + p(MA))$) at every site for each individual and used these estimates to calculate a correlation between the high coverage ancestry calls and low coverage ancestry calls. Then we calculated the percent of high confidence calls that were replicated with high confidence (>0.8 posterior) in the low coverage data for the same ancestry state ("correct") or a different ancestry state ("incorrect"). Call to the HMM for simulated low-coverage data: ancestry_hmm -e 3e-3 -a 3 0.4 0.2 0.4 -p 0 100000 0.4 -p 1 -100 0.2 -p 2 -60 0.4 –tmax 150 –tmin 2 –ne 670000. For original high coverage data we used genotype calls rather than read counts (-g) and a lower error rate (-e 1e-3).

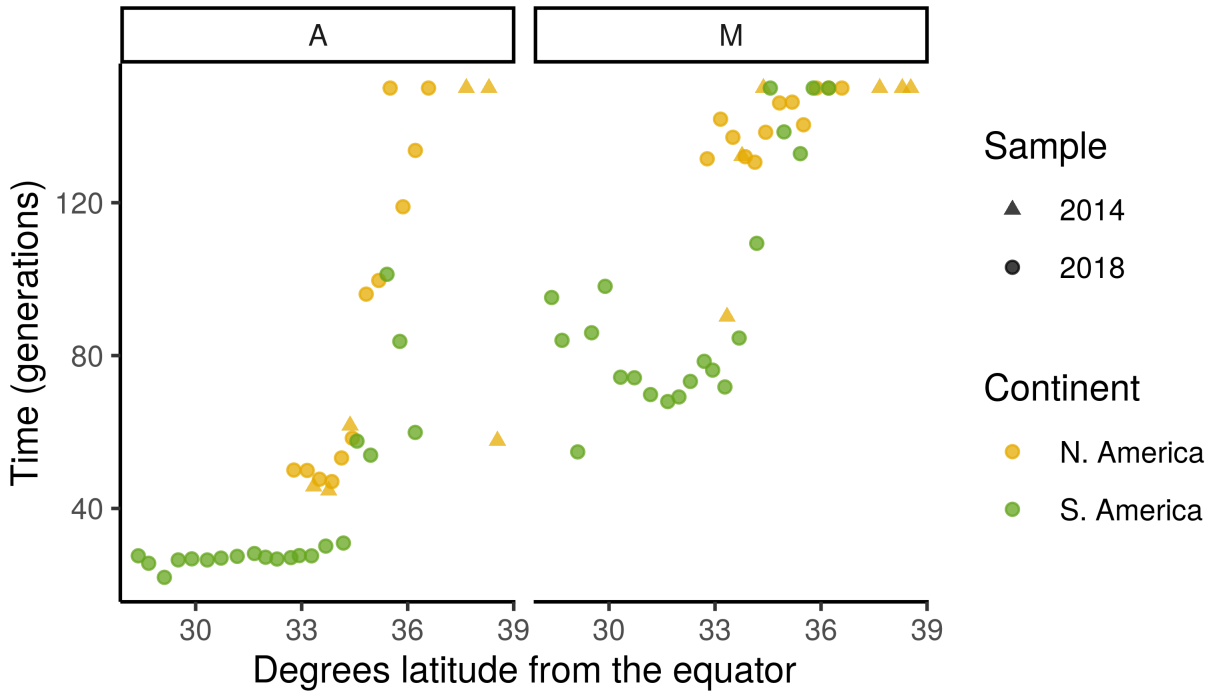


FIGURE 1.12. **Estimated generations post-admixture.** Inferred timing of migration pulses from A ancestry (left) and M ancestry (right). Each population's admixture timing is estimated separately, during local ancestry inference (ancestry_hmm), and results are plotted across latitude. We allowed a range of 2-150 generations, so the highest time estimates are truncated at 150 generations. Admixture with *scutellata* (A) ancestry began in 1956, 62 years before sampling in 2018. We have little prior information about the timing of M into C admixture, which likely varies across the Americas, but in general should pre-date admixture with A. The number of generations per year for feral honey bee populations is uncertain.

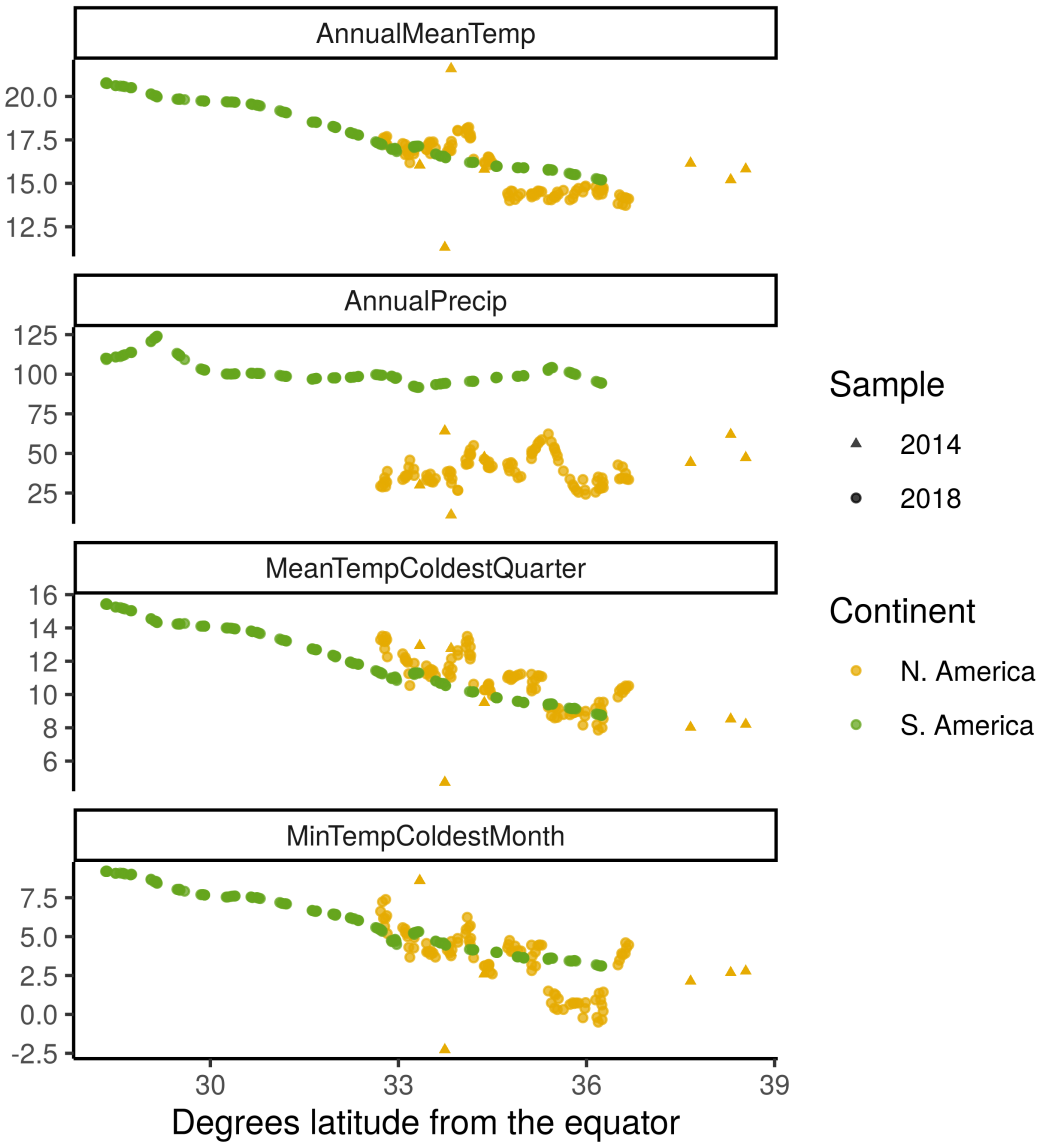


FIGURE 1.13. **Climate variables across latitude.** Bioclim climate variables for all sample sites plotted against latitude: (A) Mean annual temperature (B) Mean annual precipitation (C) Mean temperature coldest quarter (D) Minimum temperature coldest month. Two adjacent climate outliers in the N. American sample can be seen in the top two panels and represent bees from an inland desert (hot and dry) and a high altitude sampling site (cold and wet) at similar latitudes in Riverside County, CA. Bees from this same high altitude site are also outliers in the bottom two panels, having the coldest mean and minimum winter temperatures of all sites.

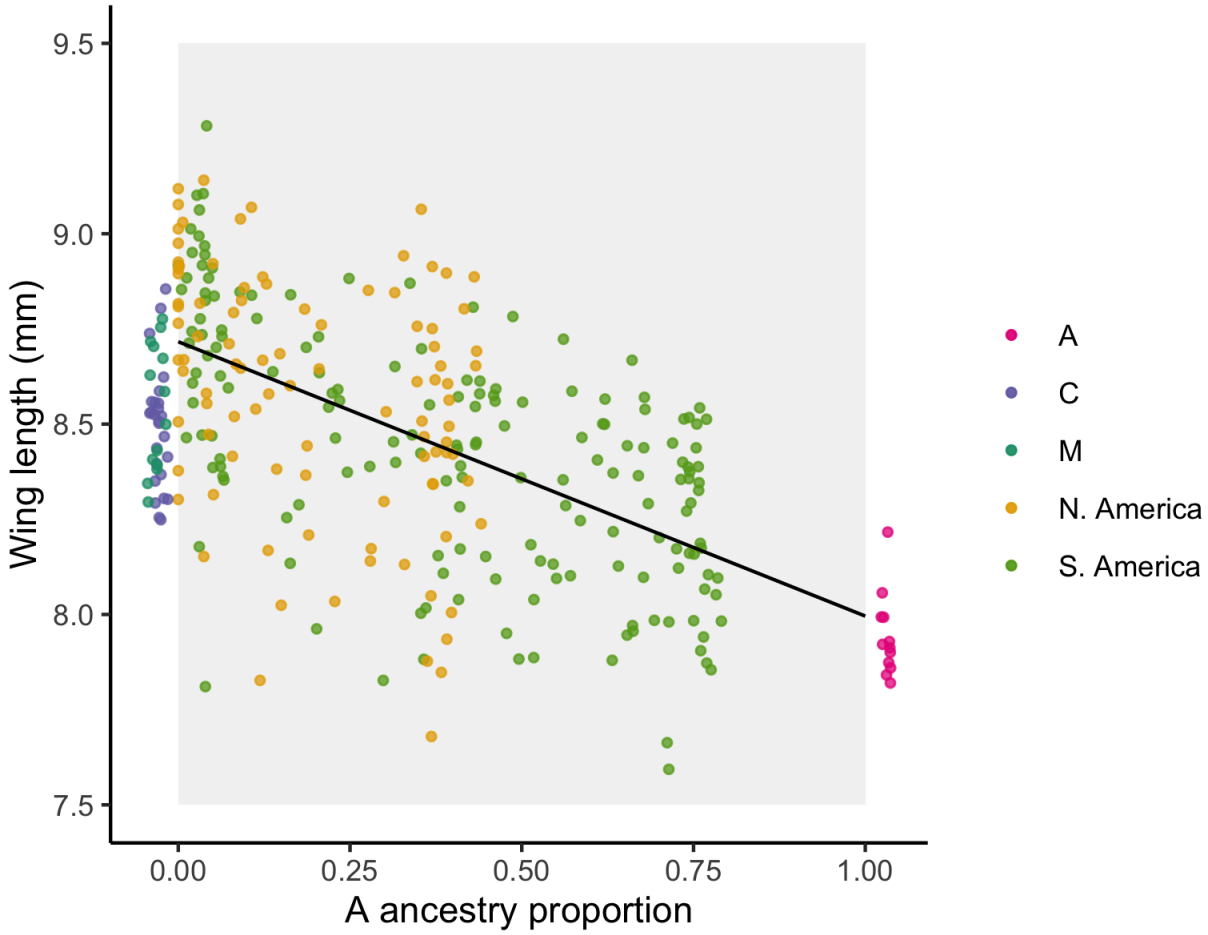


FIGURE 1.14. **Wing length predicted by ancestry.** Individual honey bees are represented as points, with wing lengths plotted along the x-axis and genomewide A ancestry proportions (NGSAdmix results) along the y-axis. We draw the best-fit regression line (slope = -0.72 mm, $F(1, 267) = 119$, $P = 3.65 \times 10^{-23}$, $R^2 = 0.31$, $n = 269$). We also include wing lengths for A, C and M reference bees from the Oberursel Collection, which we assume have none or full A ancestry. These reference bees are plotted slightly outside the range [0,1] and with jitter to facilitate viewing individual points that would otherwise all cluster on the boundaries.

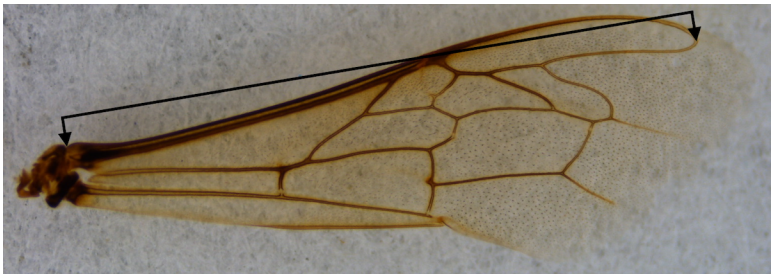


FIGURE 1.15. **Wing length measurement**. Fore wing image cropped and annotated to show length measurement taken. A full length to the tip of the wing is the standard measurement, but we use this alternative because many of our samples have significant wing tatter.

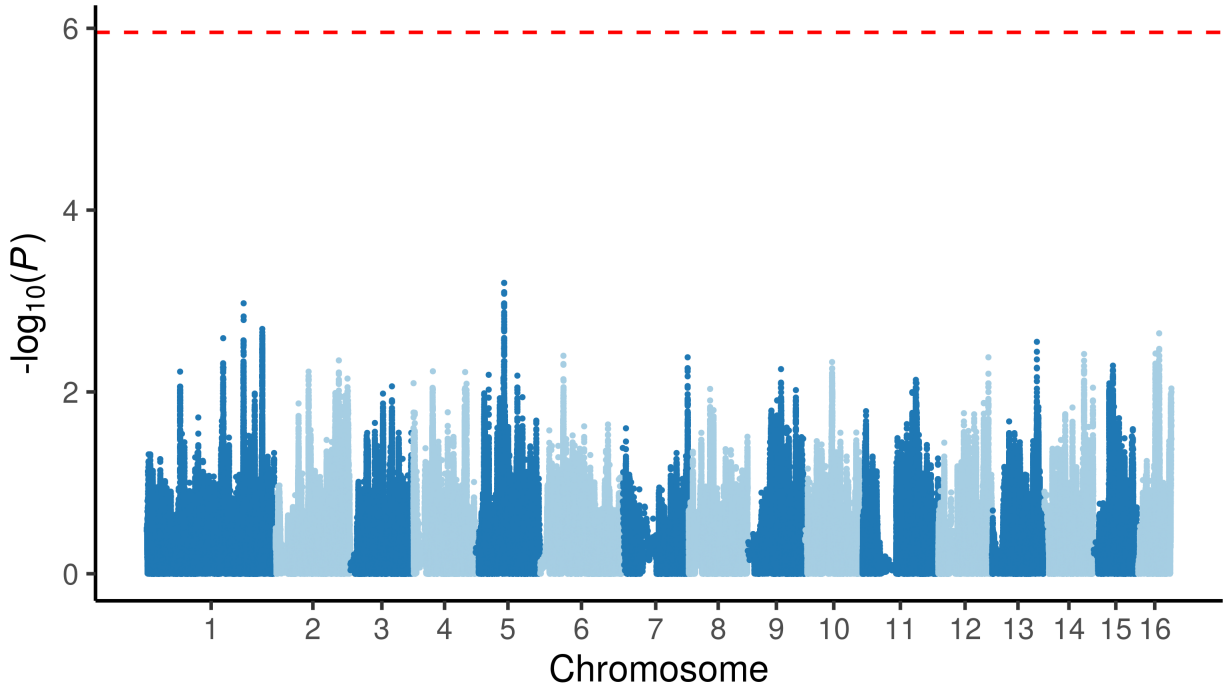


FIGURE 1.16. **Admixture mapping analysis.** We plot the p-value for each SNP across the genome, based on independent tests of association between A ancestry at that SNP and wing length. The red dashed line marks the genome-wide significance threshold for a family-wise error rate of 0.05, using a two-tailed test. In admixture mapping, SNPs are correlated, and the number of independent statistical tests depends on the number of generations recombination has had to break up ancestry blocks. Here we use an analytical approximation for the significance threshold based on 47.65 generations of admixture (population median estimate).

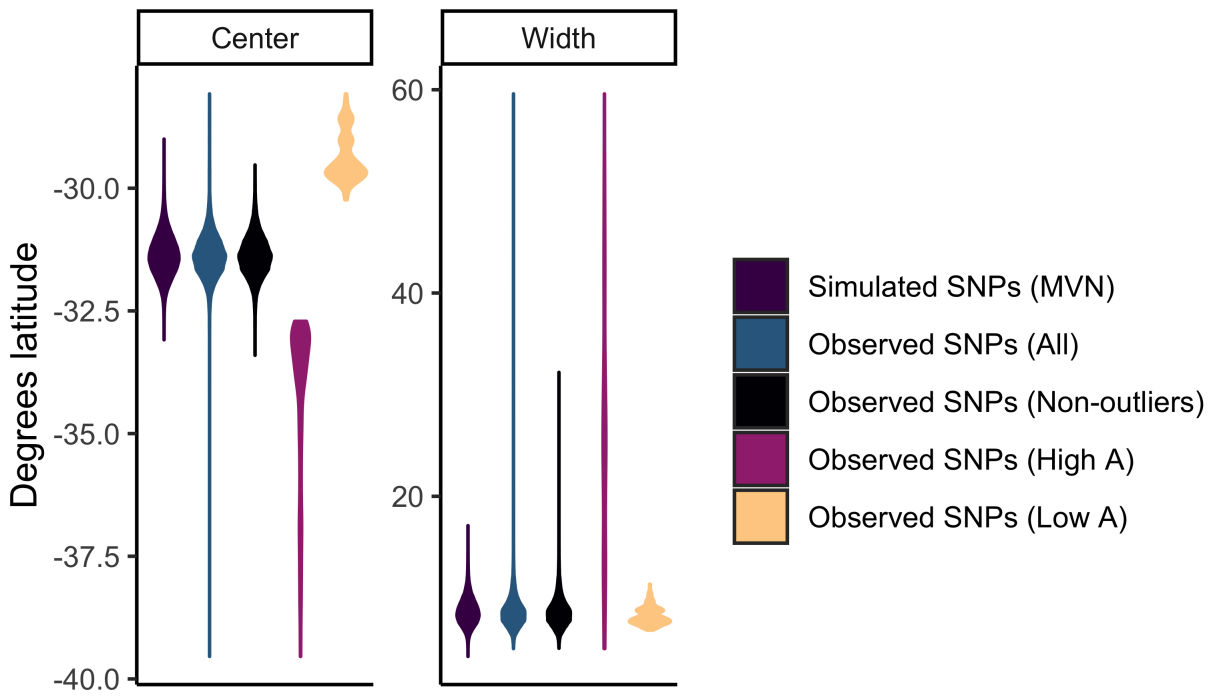


FIGURE 1.17. **Distribution of ancestry clines in South America across SNPs.** A logistic cline model was fit to observed and simulated population ancestry frequencies across latitude for S. America. Estimated cline parameters, center and width ($w = |4/b|$), are presented as violin plots. Units for both cline center and width are degrees latitude. We additionally partition observed SNPs by outlier and non-outlier status, set by 10% FDR for high or low A ancestry in South America. Individual SNP clines were only fit in South America, where we observed the full cline.

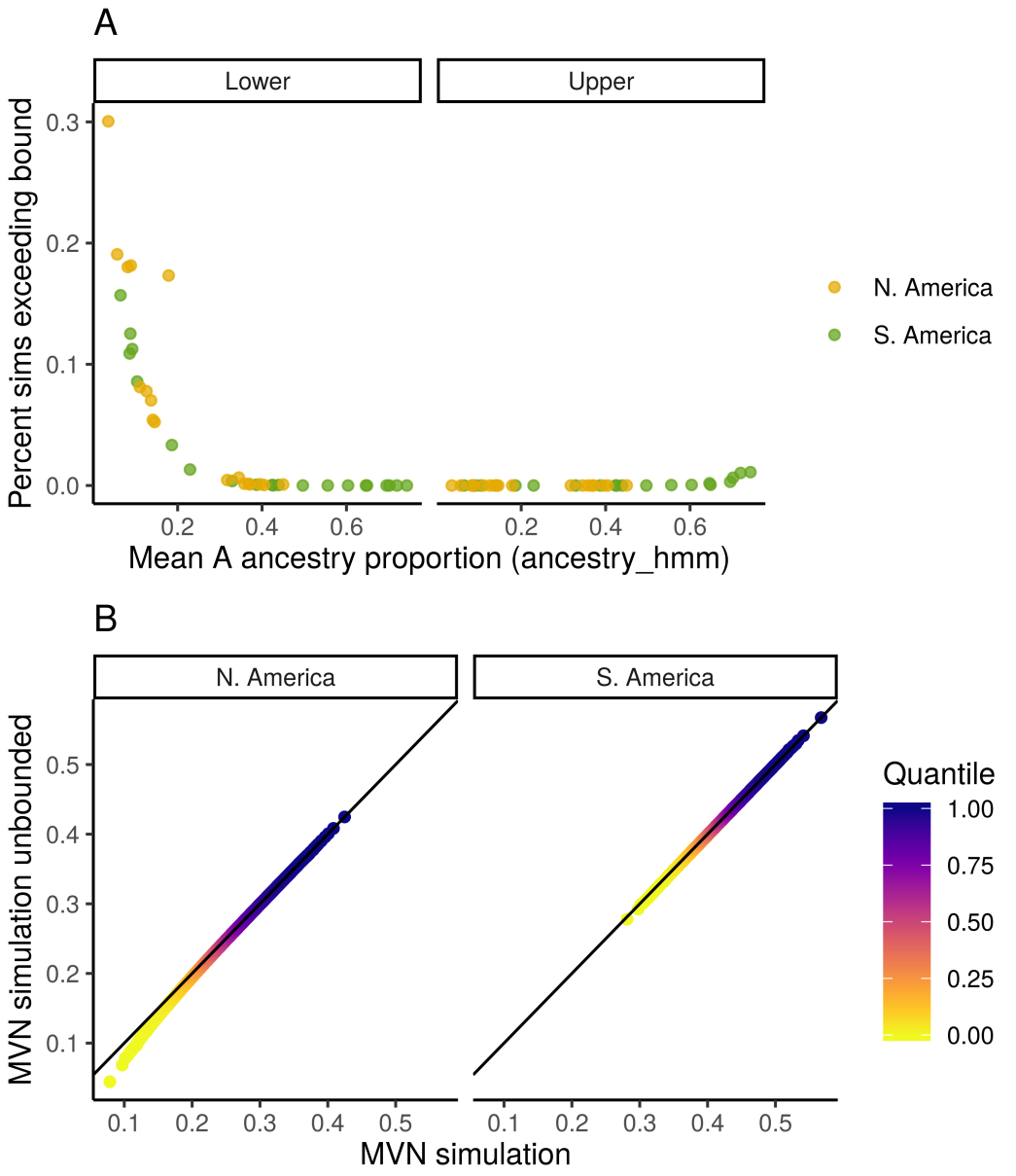


FIGURE 1.18. **Effect of truncating MVN simulated ancestry frequencies.** (A) Percent of simulated population A-ancestry frequencies exceeding lower (left) and upper (right) bounds, and thus truncated to [0,1] range. Each population is a point and the populations most affected by truncation are low-A ancestry populations with mean A-ancestry proportions close to the bound at 0. (B) QQ-plot comparing the quantiles for mean A ancestry before and after truncation in N. America (left) and S. America (right). The distribution of mean A ancestry is mostly unaffected by restricting simulated population A ancestry frequencies to the [0,1] range, but truncation does reduce model predictions of very low A-ancestry frequencies in N. America, where mean A ancestry is already low.

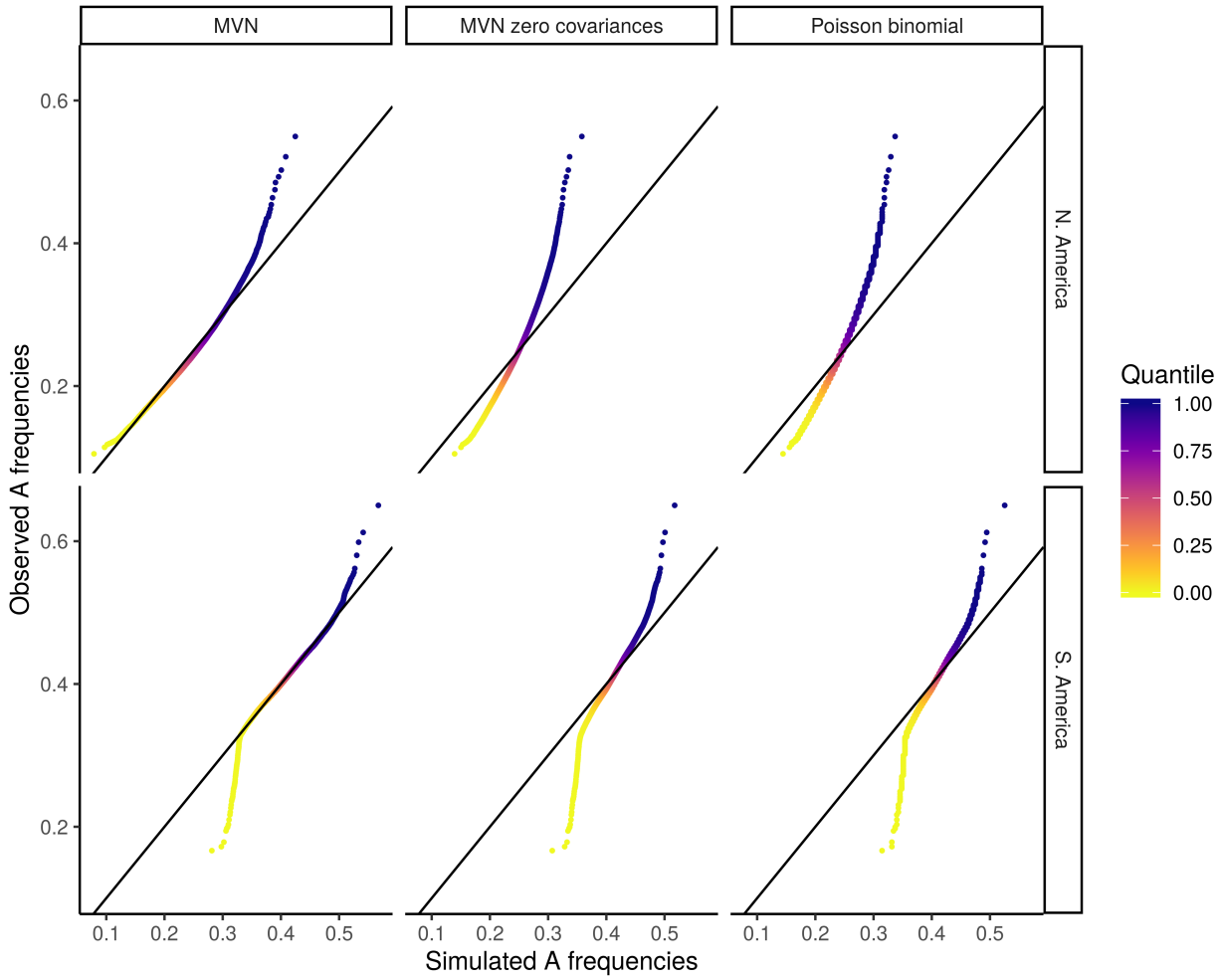


FIGURE 1.19. **Simulated vs. observed A ancestry quantiles**. QQ-plots comparing observed quantiles for mean A ancestry in North America (top) and South America (bottom) to the quantiles generated by three simulated distributions (left-to-right): MVN, MVN with zero covariances, and Poisson binomial model. Only the MVN model, allowing for covariances between populations, matches the bulk of the observed distribution.

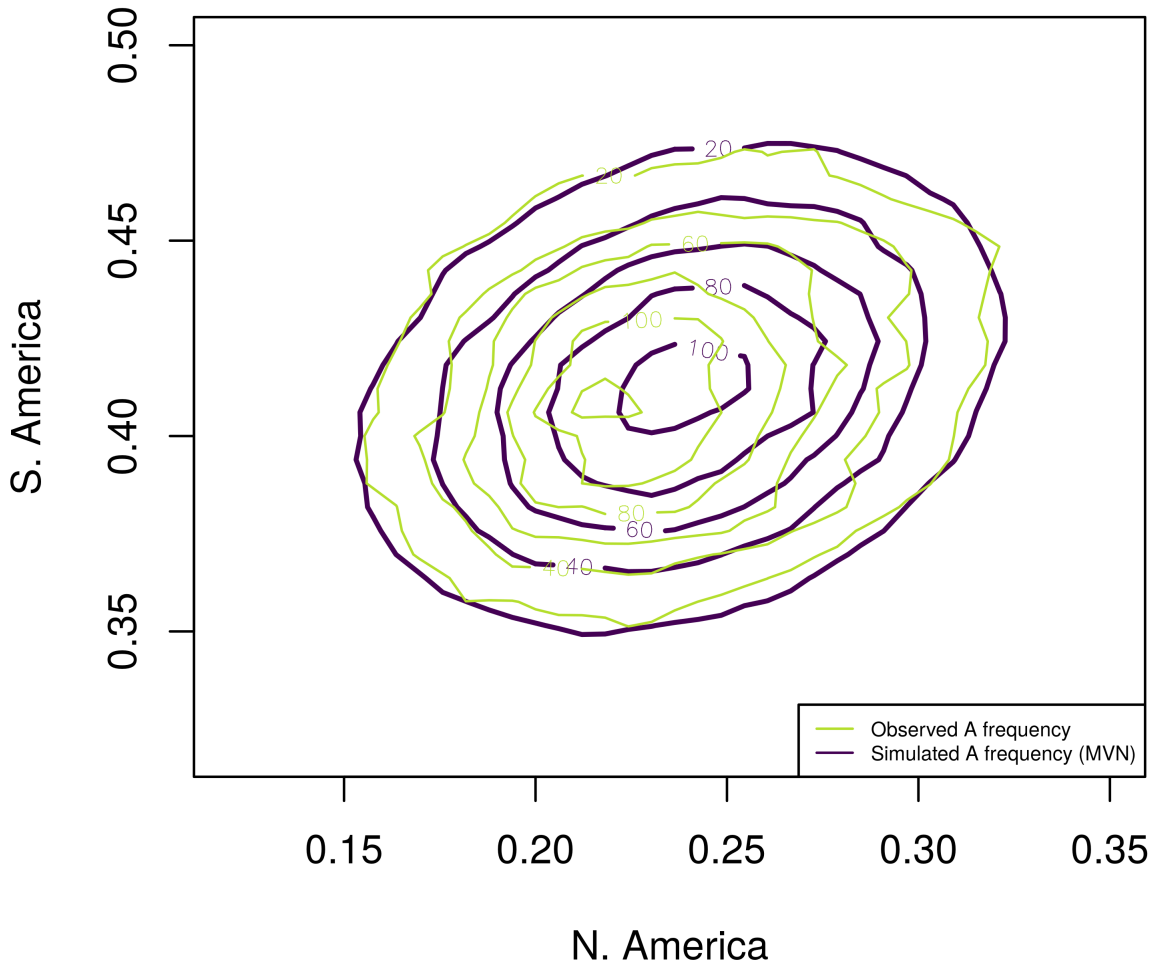


FIGURE 1.20. **Overlay of 2D density plot for observed A ancestry frequencies in North and South America compared to simulated A ancestry frequencies under a multivariate-normal model.**

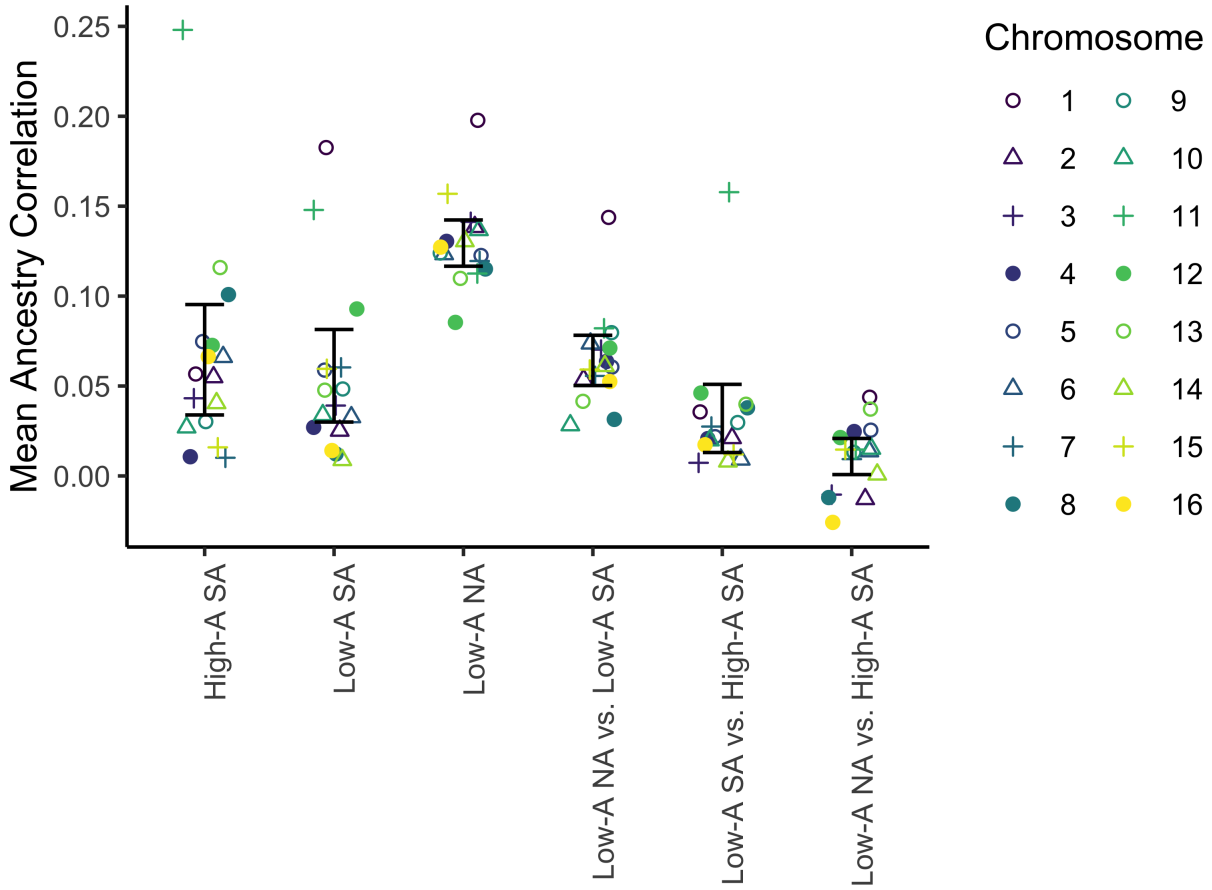


FIGURE 1.21. **Mean ancestry correlation by chromosome across populations.** Mean ancestry covariances (K matrices) were calculated separately for each chromosome, using the genome-wide mean ancestry as α, then correlations were summarised by taking the mean for each type of population comparison, within and between continents and low vs. high A regions. Error bars show the normal-approximated 95% confidence intervals around these means. We divided populations in South America into low A and high A groups relative to the cline center. Low A populations are found at higher latitudes and correspondingly cooler climates. All North American samples come from the low-A side of the cline. On average across chromosomes, low-A South American populations share higher ancestry correlations with low-A North American populations than with geographically closer high-A South American populations (0.032: $CI_{95}[0.011, .053]$, P = .005, paired 2-sided t-test). Two chromosomes harbor large outlier regions consistent with their elevated correlations shown here: Chromosome 1 has a large cluster of loci with high A ancestry in North and South America while chromosome 11 has a wide region of low A ancestry exclusive to South America. The results do not change qualitatively if these two outlier chromosomes are both removed from the analysis (0.035: $CI_{95}[0.022, 0.047]$, P = $4.8x10^{-5}$, paired 2-sided t-test).

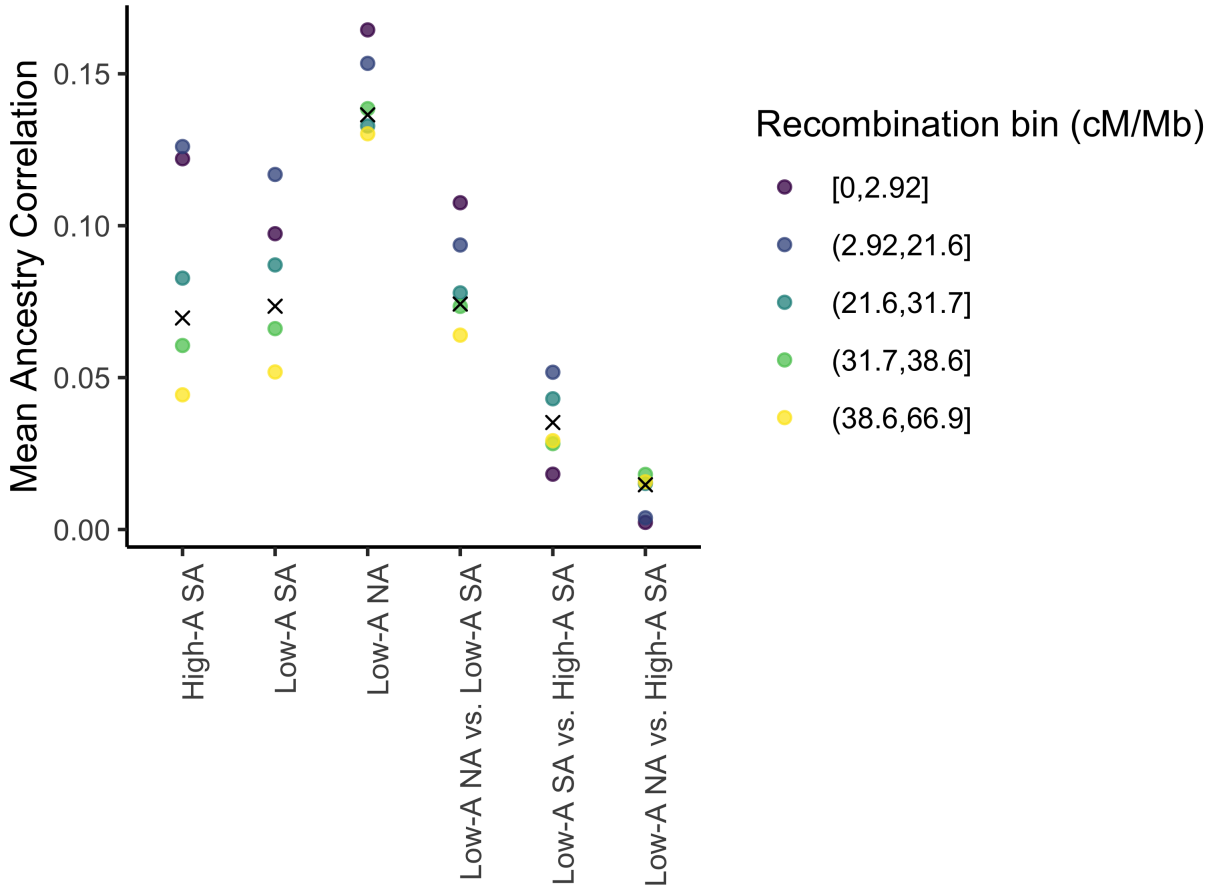


FIGURE 1.22. **Mean correlation for population pairs by recombination rate.** Mean ancestry covariances (K matrices) were calculated separately for each of the 5 recombination rate quintiles, using the genome-wide mean ancestry as α, then correlations were summarised by taking the mean for each type of population comparison, within and between continents and low vs. high A regions. About half of the South American populations, and all of the sampled North American populations come from the low-A side of the hybrid zone (relative to the estimated cline center). The genomewide mean is additionally shown as an X. On average across recombination bins, low-A South American populations share higher ancestry correlations with low-A North American populations than with geographically closer high-A South American populations (0.049: $CI_{95}[0.021, .078]$, P = .009, paired 2-sided t-test)

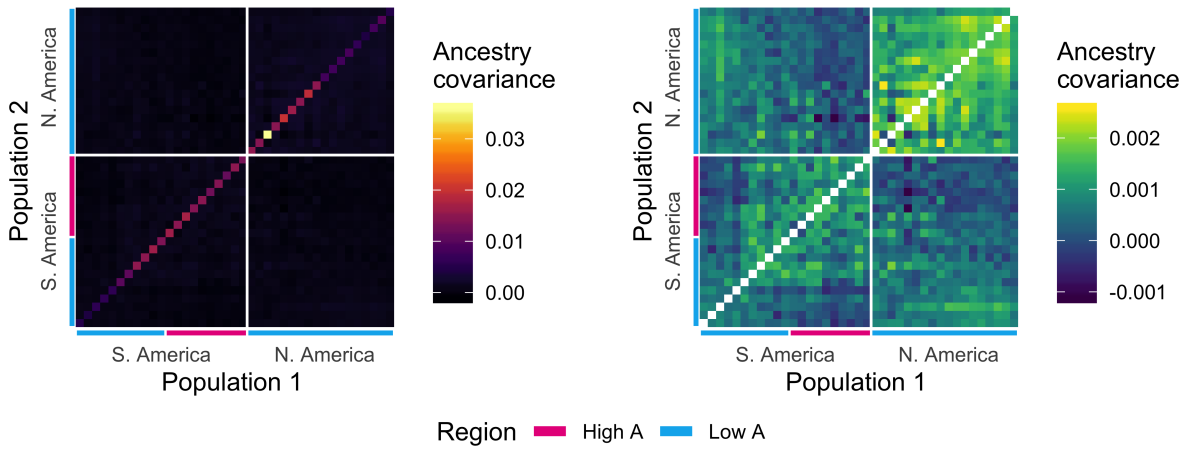


FIGURE 1.23. **Ancestry covariances across populations.** Ancestry covariance matrix (see methods). Populations are ordered by latitude, with high and low A sides of each hybrid zone defined relative to the estimated genomewide cline center. Within-population variances (left) are shown separately from between-population covariances (right) because of the drastically different scales. Populations near the cline center have higher ancestry variances (most clearly seen in the diagonal elements) because they have A ancestry proportions closer to 50%. Drift and finite sample sizes also contribute to the observed ancestry variances. The third lowest latitude population in the North American cline with exceptionally high ancestry variance is Avalon, sampled from Catalina Island off the coast of California. The observed excess covariance between the two distant ends of these clines is unexpected under a simple model of spread North and South out of Brazil.

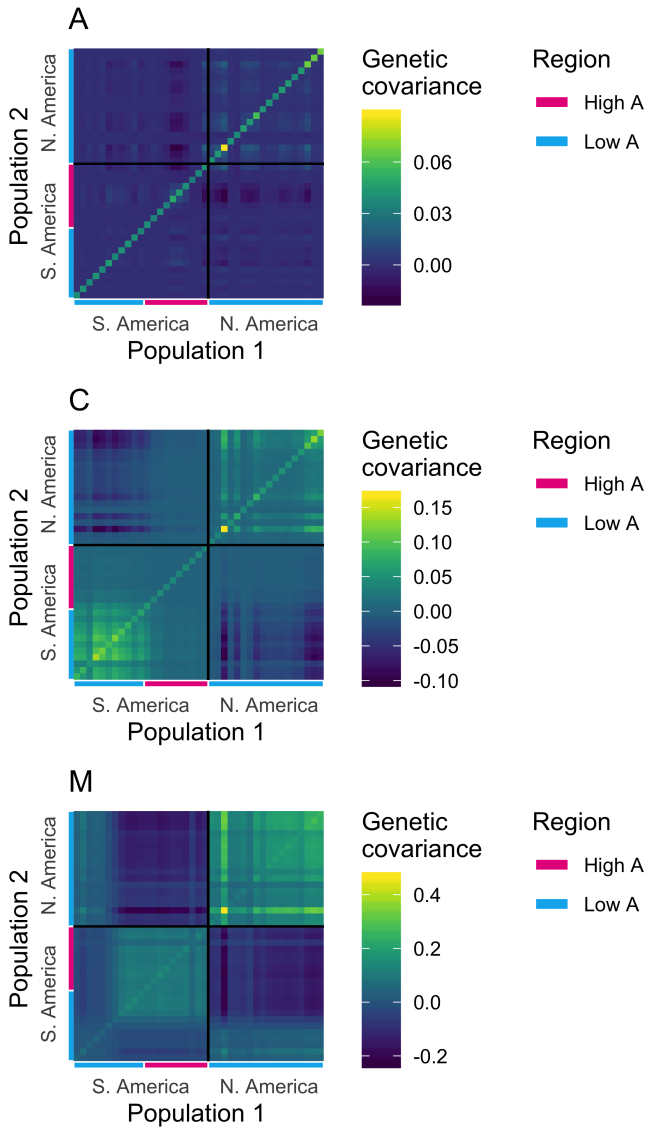


FIGURE 1.24. **Genetic covariance within ancestry.** Genetic covariances within A (top), C (middle) and M (bottom) ancestry. Colors represent the population mean genetic covariance between individuals, and the range of values varies by ancestry (note: color bars have different scales). While kinship creates strictly positive covariances, here we observe some negative values because we can only calculate co-variation around the empirical mean combined sample allele frequency, not the true ancestral allele frequency (which is unknown). Population mean covariances were summarised from an individual-by-individual covariance matrix generated using PCAngsd from bam files filtered to only include regions of the genome with high confidence homozygous ancestry calls for the focal ancestry (posterior >0.8 from ancestry_hmm). High and low A sides of each hybrid zone are defined relative to the estimated genomewide cline center

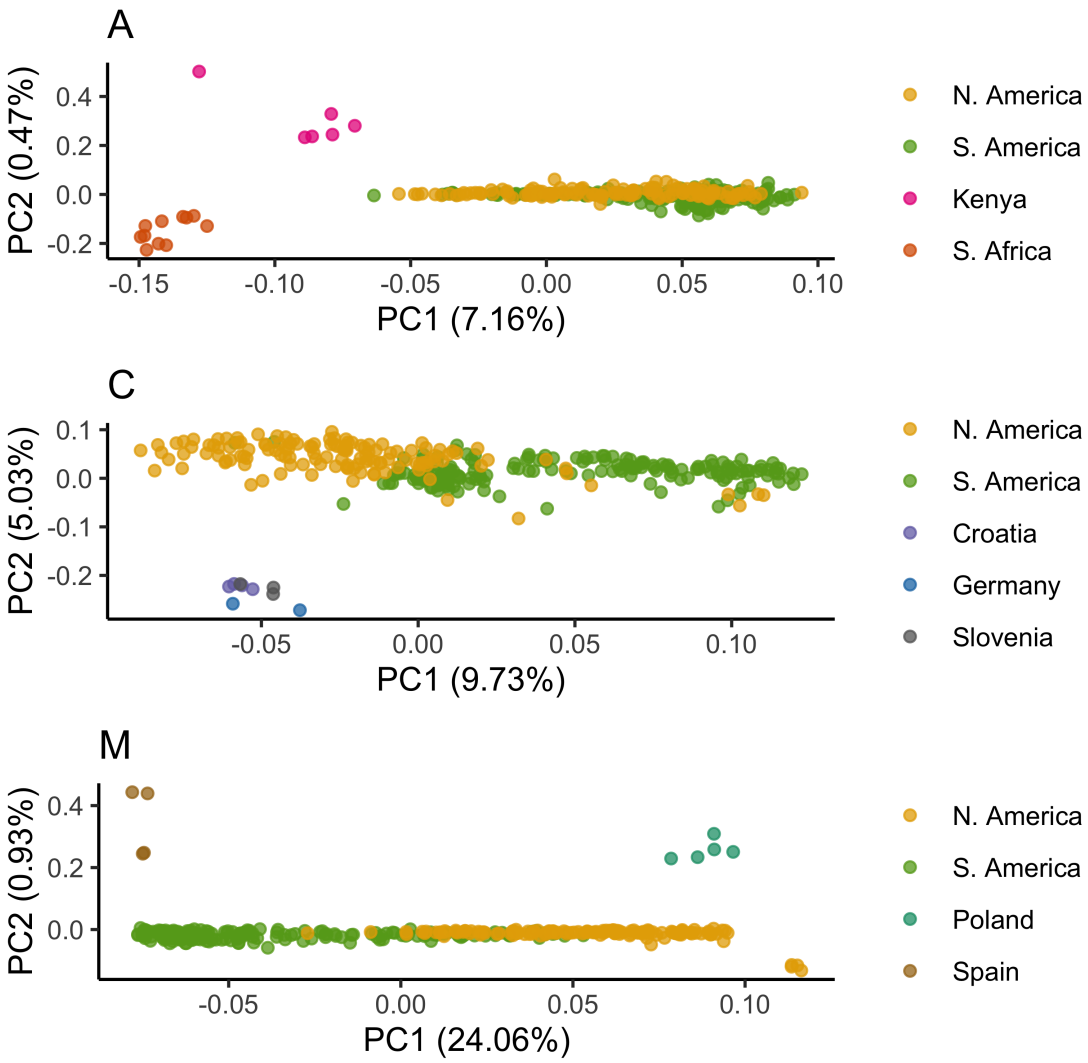


FIGURE 1.25. **Principle components analysis of genetic variation within ancestry.** PCA analysis of A (top), C (middle) and M (bottom) ancestry. Analysis was performed using PCAngsd using all reference samples of the focal ancestry and sequence data from the hybrid zones filtered to only include regions of the genome with high confidence homozygous ancestry calls for the focal ancestry (posterior >0.8 from ancestry_hmm). Each bee is a point, colored by sample location. The two hybrid zones form somewhat separable clusters for European (C and M) ancestry, but not *scutellata* (A) ancestry. The major axis of genetic diversity within M ancestry in the Americas (PC1) mirrors pre-existing population structure within Europe between *Apis mellifera mellifera* (Poland) and *Apis mellifera iberiensis* (Spain), two well-known honeybee subspecies that may have different historical import rates to different regions.

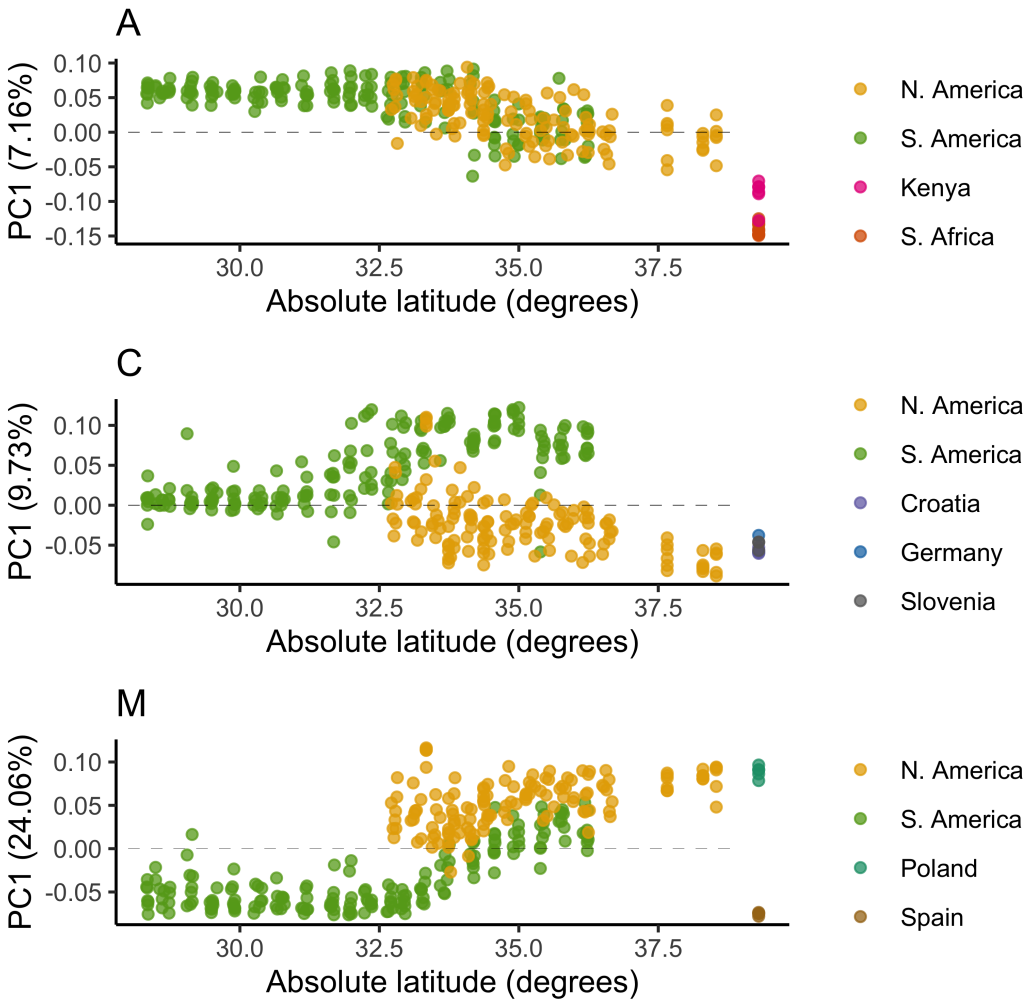


FIGURE 1.26. **Genetic variation within ancestry by latitude.** Here we plot the first principal component for genetic diversity within A (top), C (middle) and M (bottom) ancestry against absolute latitude of sampling location within the hybrid zone (see Fig 1.25 for original PCA). Each bee is a point and reference bees are plotted to the side (not at their actual latitude). Bees are colored by sample location. Note that bees with very low amounts of the focal ancestry (higher latitudes A ancestry or lower latitudes C ancestry) fall close to zero on PC1 (dashed line), which, despite using a method designed to account for low coverage data (PCAngsd), may simply be an artifact of low information per individual bee for genetic diversity within a low-frequency ancestry. A ancestry shows very little population structure along PC1 by continent or latitude. C is the dominant ancestry at higher latitudes in both zones and shows greater separation between the two ends of the zones (higher absolute latitude for both) than within South America. M ancestry at lower latitudes in South America is more similar to *Apis mellifera iberiensis* (Spain) than M ancestry elsewhere in the Americas.

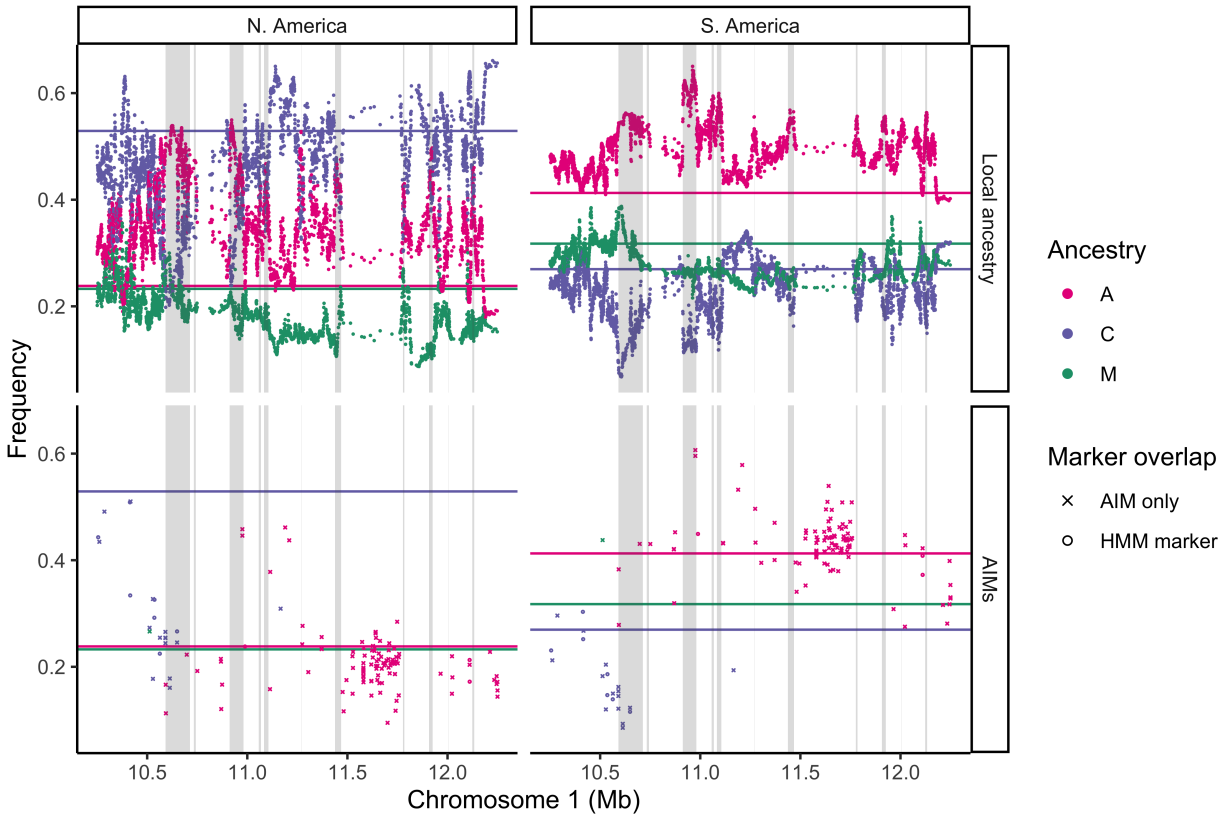


FIGURE 1.27. **Ancestry and AIM frequencies for high shared A outliers on chr1.** Zoomed-in view of the region on chromosome 1 with a cluster of high A ancestry peaks in both North America (left) and South America (right), with shared outlier regions meeting a 10% FDR for high A ancestry on both continents shaded in grey. (Top) *Scutellata* (A), western European (M) and eastern European (C) local ancestry estimates at each HMM marker. (Bottom) Mean frequency of ancestry informative markers AIMs (see methods), most of which were not included in the ancestry_hmm inference ('AIM only') due to thinning.

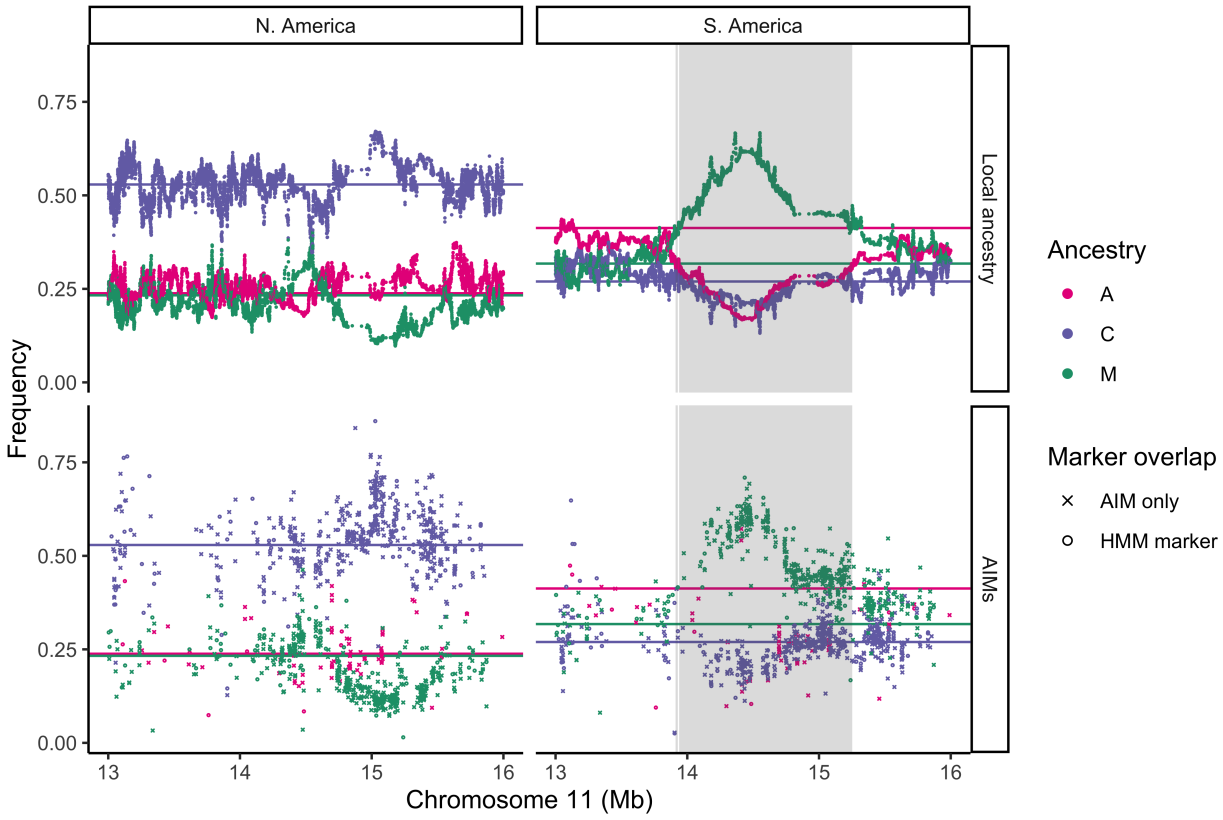


FIGURE 1.28. **Ancestry and AIM frequencies for low A outlier region on chr11.** Zoomed-in view of the 1.4Mb region on chromosome 11 with high western European ancestry (M) in South America (right) but not in North America (left), with the outlier region meeting 10% FDR highlighted in grey. (Top) *Scutellata* (A), western European (M) and eastern European (C) local ancestry estimates at each HMM marker. (Bottom) Mean frequency of ancestry informative markers AIMs (see methods), most of which were not included in the ancestry_hmm inference ('AIM only') due to thinning.

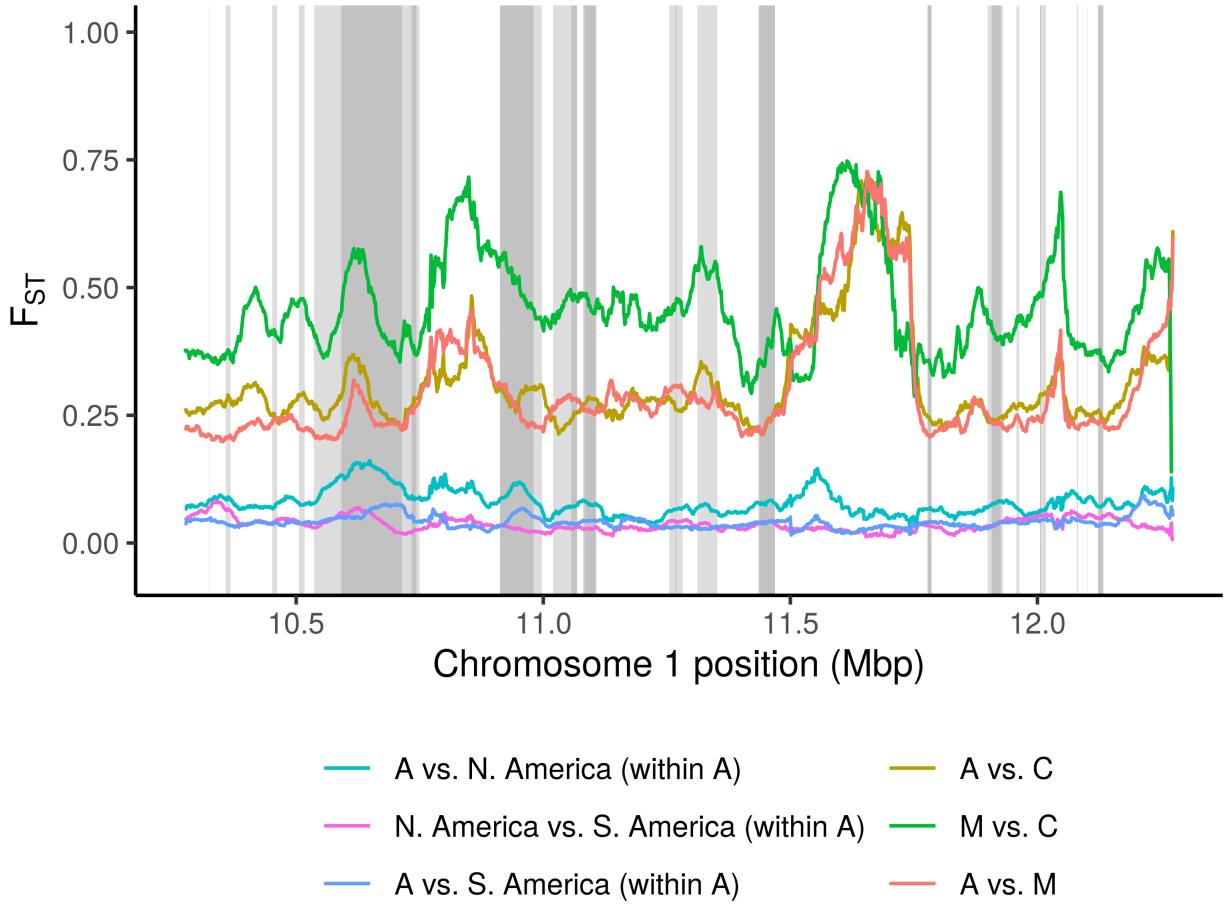


FIGURE 1.29. **Differentiation across shared high A outliers on chr1.** F_{ST} across the region on chromosome 1 with shared high A ancestry outliers. Outlier regions meeting a 10% FDR in both hybrid zones are highlighted in darker grey, while those meeting a 10% FDR in only one hybrid zone are highlighted in lighter grey. Per-SNP F_{ST} is averaged within sliding 50kb windows. In addition to the three ancestry reference panels (A, C, & M), we include contrasts for the subset of individuals in each hybrid zone with high-confidence homozygous A ancestry.

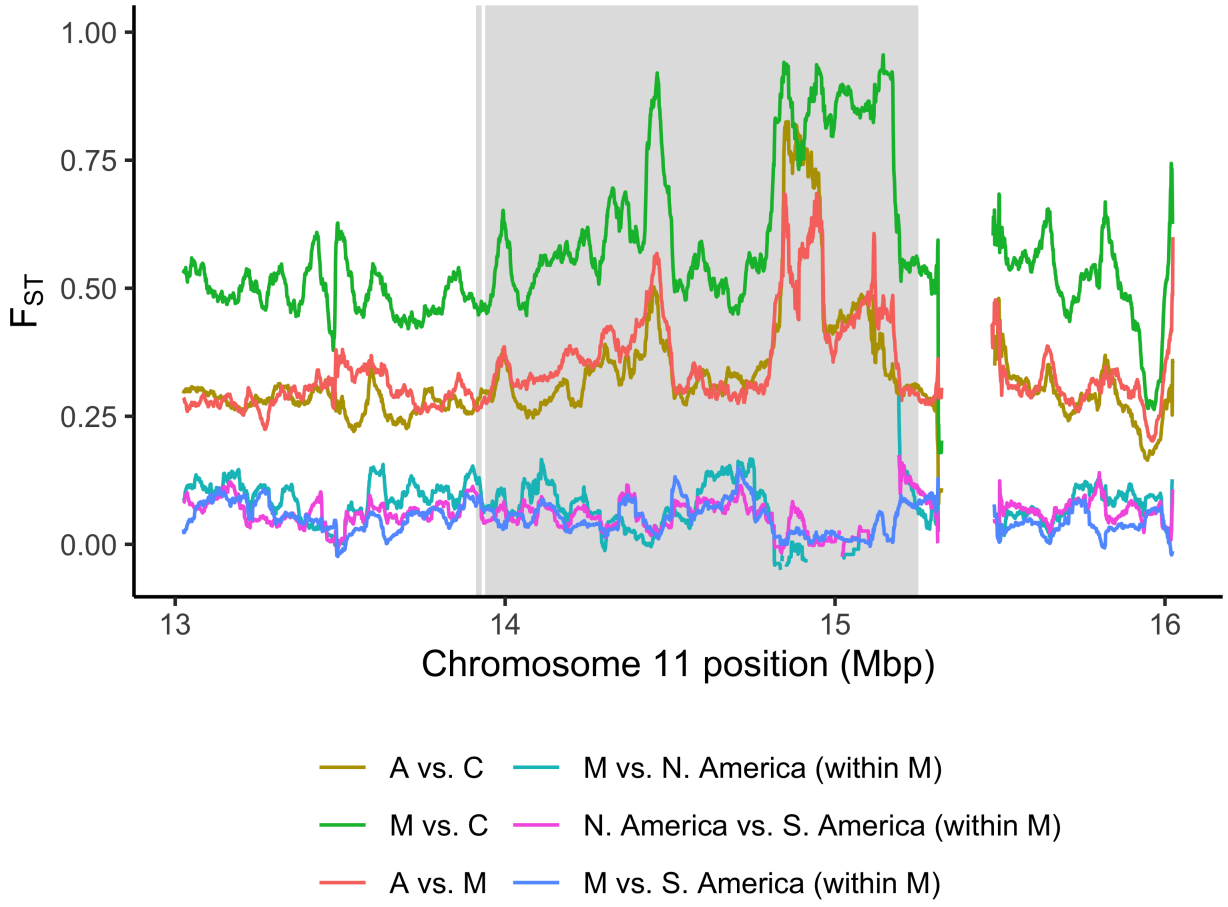


FIGURE 1.30. **Differentiation across shared high A outliers on chr11.** F_{ST} across the region on chromosome 11 with shared high A ancestry outliers. Outlier regions meeting a 10% FDR in both hybrid zones are highlighted in darker grey, while those meeting a 10% FDR in only one hybrid zone are highlighted in lighter grey. Per-SNP F_{ST} is averaged within sliding 50kb windows. In addition to the three ancestry reference panels (A, C, & M), we include contrasts for the subset of individuals in each hybrid zone with high-confidence homozygous A ancestry.

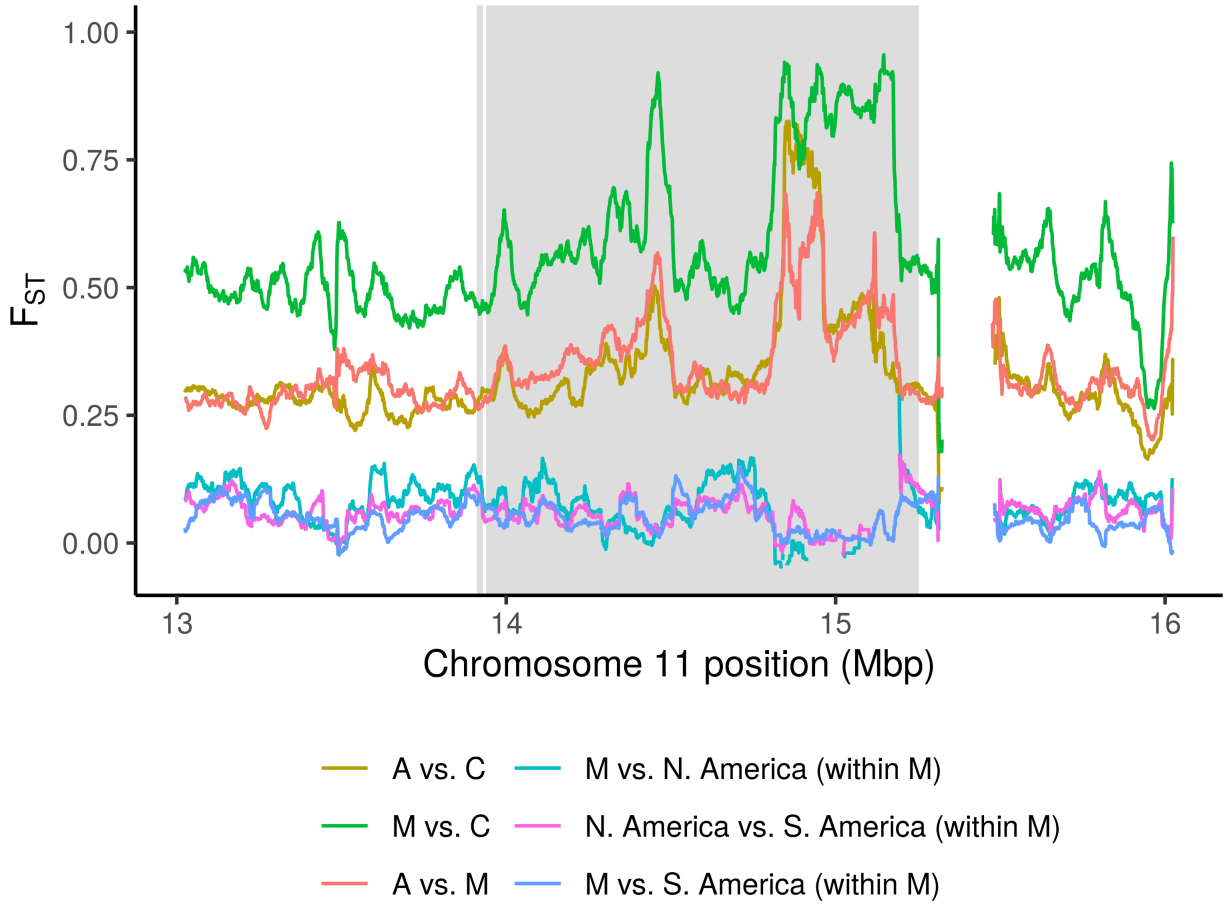


FIGURE 1.31. **Differentiation across low A outlier region on chr11.** F_{ST} across the 1.4Mb region on chromosome 11 with high western European ancestry (M) in South America but not in North America (left), with the outlier region meeting 10% FDR highlighted in grey. Per-SNP F_{ST} is averaged within sliding 50kb windows. In addition to the three ancestry reference panels (A, C, & M), we include contrasts for the subset of individuals in each hybrid zone with high-confidence homozygous M ancestry. Windows are dropped if fewer than 10 SNPs have 2 individuals with data, which produces gaps in the contrasts with N. American M ancestry because this hybrid zone does not have elevated M ancestry in this region and at many SNPs very few individuals have high-confidence homozygous M ancestry. Top peaks in M vs. C, M vs. A, and C vs. A contrasts seen within this region reach the 99.5, 98.0, and 99.8 percentiles (respectively) for 50kb windows genome-wide.

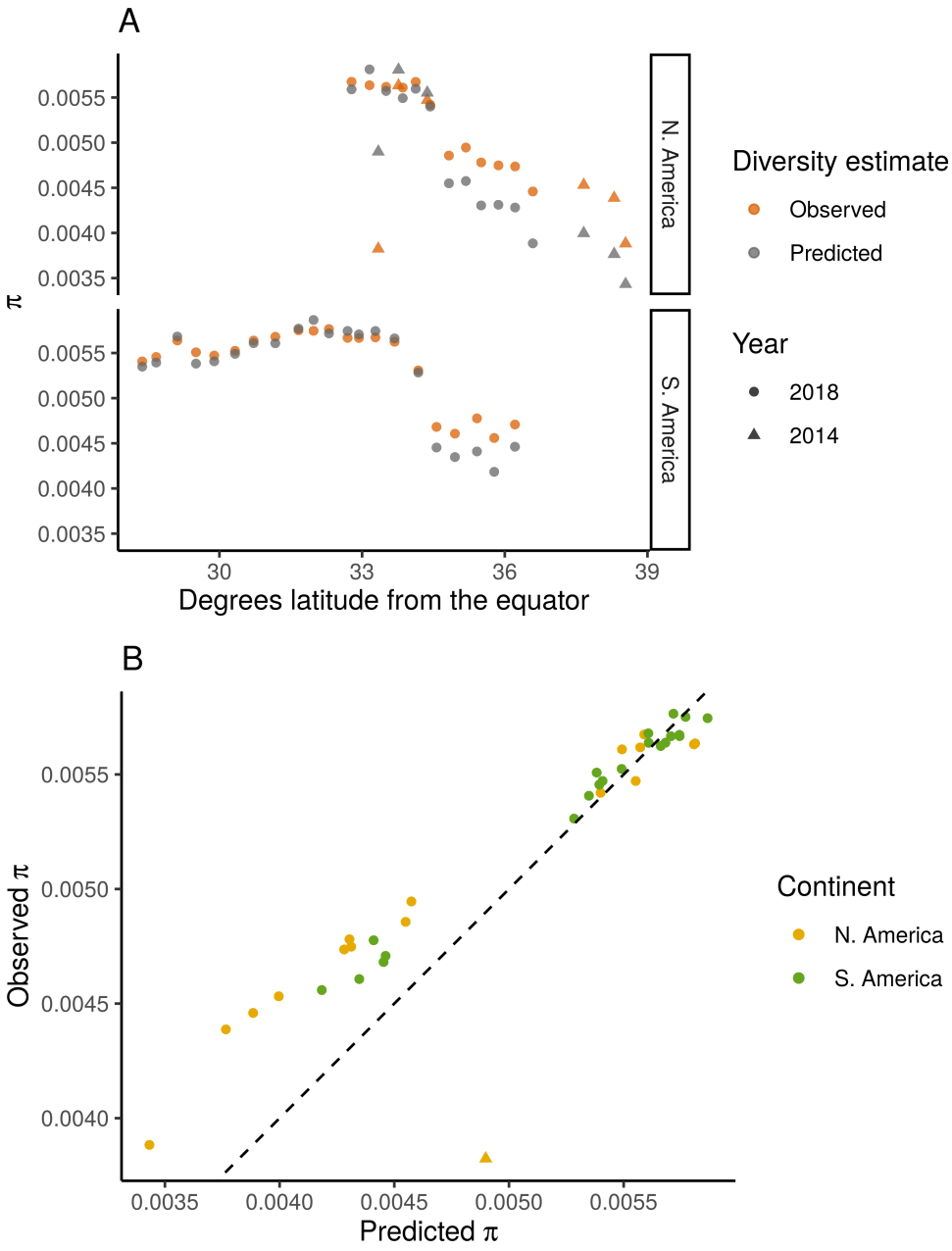


FIGURE 1.32. **Comparison of observed and predicted diversity.** (A) Observed and predicted allelic diversity (π) for each population across latitude. To predict π for a specific population, we calculated the expected allele frequency based on a mixture of A, C, and M reference population allele frequencies, weighted by the population's estimated admixture fractions of these three ancestries. (B) Plot of predicted vs. observed π for each population for direct comparison to the 1-to-1 line (dashed). Avalon (California 2014, marked as a triangle) is a clear outlier, with low diversity for its admixture fraction.

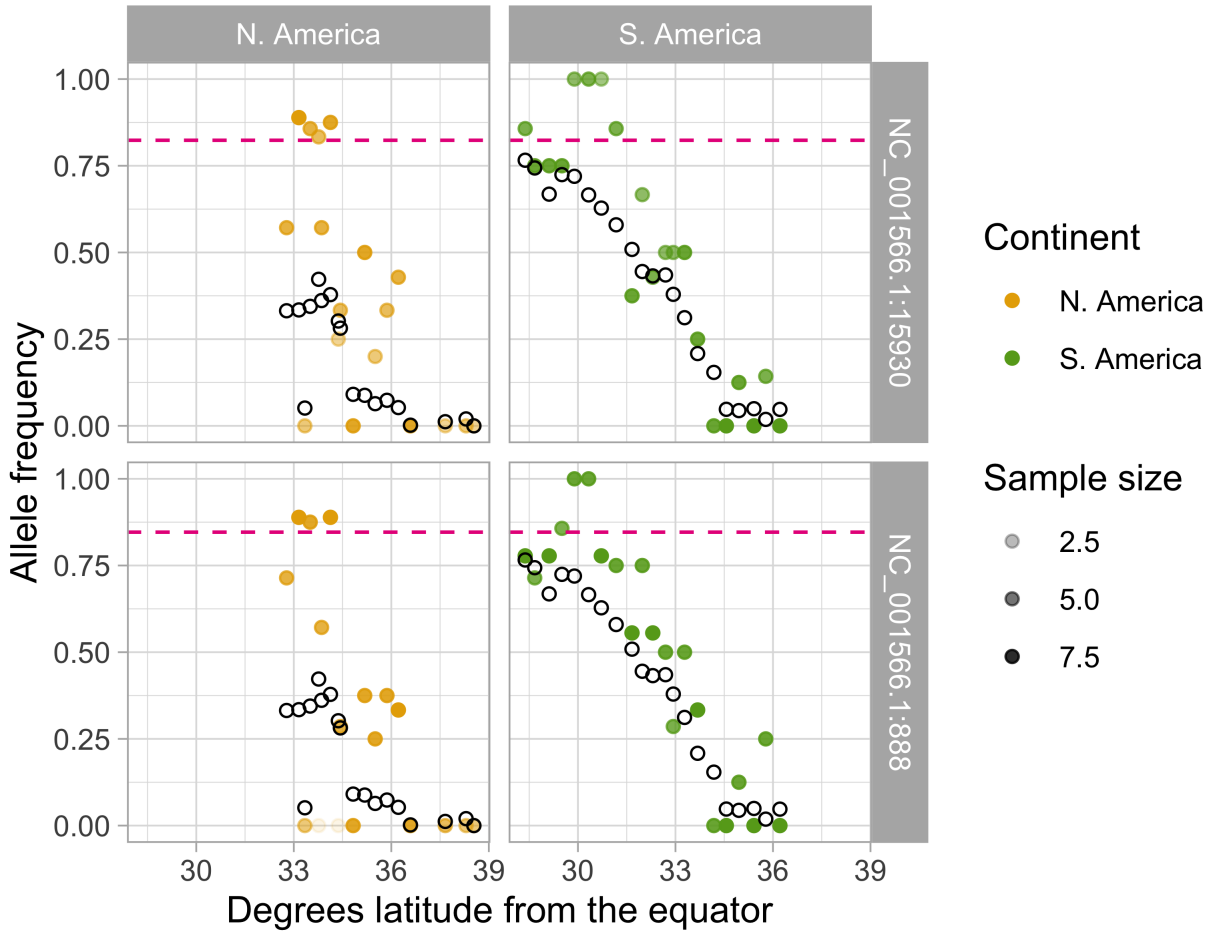


FIGURE 1.33. **Mitochrondrial clines.** Out of 82 SNPs on the mitochrondria, we identified two with more than 80% estimated frequency difference between *scutellata* (A) and European (C & M) reference panels. Estimated allele frequencies at these SNPs for each population in North America (left) and South America (right) are plotted in color. For comparison, population mean genomewide A ancestry proportions (NGSAdmix) are plotted as open black circles. At both SNPs, estimated M and C allele frequencies are zero (not shown) and estimated A allele frequencies are high but not at fixation (plotted as pink dashed lines). Bees sequenced in this study have low coverage across most of the mtDNA sequence, which prevented us from constructing a phylogenetic tree for full mitochondrial haplotypes and creates uncertainty in our allele frequency estimates here. To reflect this uncertainty, points are shaded by the sample size (i.e. the number of mtDNA haplotypes in a population for which we were able to call a consensus base).

CHAPTER 2

Selective sorting of ancestral introgression in maize and teosinte along an elevational cline

Erin Calfee[1,2], Daniel Gates[2,7], Anne Lorant[3,5], M. Taylor Perkins[2,6], Graham Coop[1,2,†], and Jeffrey Ross-Ibarra[1,2,4,†]

[1] Center for Population Biology, University of California, Davis, United States of America

[2] Department of Evolution and Ecology, University of California, Davis, United States of America

[3] Department of Plant Sciences, University of California, Davis, United States of America

[4] Genome Center, University of California, Davis, United States of America

[5] Laboratoire de Biologie Moléculaire et Cellulaire du Cancer, Hôpital Kirchberg, Luxembourg (current address)

[6] Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, United States of America (current address)

[7] Checkerspot, Inc., Berkeley, California, United States of America (current address)

† co-mentors.

Abstract

While often deleterious, hybridization can also be a key source of genetic variation and pre-adapted haplotypes, enabling rapid evolution and niche expansion. Here we evaluate these opposing selection forces on introgressed ancestry between maize (*Zea mays* ssp. *mays*) and its wild teosinte relative, *mexicana* (*Zea mays* ssp. *mexicana*). Introgression from ecologically diverse teosinte may have facilitated maize's global range expansion, in particular to challenging high elevation regions ($>$ 1500 m). We generated low-coverage genome sequencing data for 348 maize and *mexicana* individuals to evaluate patterns of introgression in 14 sympatric population pairs, spanning the

elevational range of *mexicana*, a teosinte endemic to the mountains of Mexico. While recent hybrids are commonly observed in sympatric populations and *mexicana* demonstrates fine-scale local adaptation, we find that the majority of *mexicana* ancestry tracts introgressed >1000 generations ago.

This *mexicana* ancestry seems to have maintained much of its diversity and likely came from a common ancestral source, rather than contemporary sympatric populations, resulting in relatively low F_{ST} between *mexicana* ancestry tracts sampled from geographically distant maize populations.

Introgressed *mexicana* ancestry is reduced in lower-recombination rate quintiles of the genome and around domestication genes, consistent with pervasive selection against introgression. However, we also find *mexicana* ancestry increases across the sampled elevational gradient and that high introgression peaks are most commonly shared among high-elevation maize populations, consistent with introgression from *mexicana* facilitating adaptation to the highland environment. In the other direction, we find patterns consistent with adaptive and clinal introgression of maize ancestry into sympatric *mexicana* at many loci across the genome, suggesting that maize also contributes to adaptation in *mexicana*, especially at the lower end of its elevational range. In sympatric maize, in addition to high introgression regions we find many genomic regions where selection for local adaptation maintains steep gradients in introgressed *mexicana* ancestry across elevation, including at least two inversions: the well-characterized 14 Mb *Inv4m* on chromosome 4 and a new 3 Mb inversion *Inv9f* surrounding the *macrohairless1* locus on chromosome 9. Most outlier loci with high *mexicana* introgression show no signals of sweeps or local sourcing from sympatric populations and so likely represent ancestral introgression sorted by selection, resulting in correlated but distinct outcomes of introgression in different contemporary maize landrace populations.

Introduction

Interbreeding between partially diverged species or subspecies can result in admixed individuals with low fitness, e.g. due to hybrid incompatibilities [**118**,**119**,**120**]. Consistent with the view that hybridization is often deleterious, a growing number of species show evidence of pervasive selection against introgressed ancestry [**121**, **122**, **123**, **124**, **125**, **126**, **127**, **128**, **129**, **130**]. At the same time, introgression can be a source of novel genetic variation and efficiently introduce haplotypes

carrying sets of locally adapted alleles, with the potential for rapid adaptation to new ecological challenges [**131**]. Indeed, admixture has been linked to adaptive species radiations and/or rapid niche expansions in a number of natural systems, including mosquitoes [**132**], *Drosophila* [**133**], butterflies [**126**], cichlids [**134**], sunflowers [**135**], wild tomatoes [**136**] and yeast [**137**, **138**]. In addition, introgression from wild relatives has facilitated the broad range expansions of multiple domesticated crops (reviewed in [**139**] and [**140**]), and gene flow from crops back into their wild relatives has in some cases opened up novel 'weedy' niches [**141**].

Maize (*Zea mays* ssp. *mays*) is an ideal system to study selection on admixed ancestry and the effects on range expansion, as it has colonized nearly every human-inhabited ecosystem around the world [**142**] and interbreeds with a number of wild relatives genetically adapted to distinct ecologies [**143**, **144**]. In Mexico, highland maize represents an early major niche expansion that may have been facilitated by introgression. Approximately 9 thousand years ago, maize (*Zea mays* ssp. *mays*) was domesticated in the Balsas River Valley in Mexico from a lowland-adapted subspecies of teosinte (*Zea mays* ssp. *parviglumis* [**145**]), which grows readily at sea level and up to about 2000 meters [**146**]. In contrast, *Zea mays* ssp. *mexicana*, which diverged from *parviglumis* about 60 thousand years ago [**147**], is endemic to highland regions in Mexico (∼1500-3000 meters in elevation) where it has adapted to a number of ecological challenges: a cooler, drier climate with higher UV intensity, different soil nutrient composition, and a shorter growing season necessitating earlier flowering times [**148**, **149**, **150**, **151**, **152**].

Maize was introduced as a crop to the mountains of Mexico around 6.2 thousand years ago [**153**], and it is thought that gene flow from *mexicana* assisted in adaptation to high elevation selection pressures. Highland maize and *mexicana* share a number of putatively adaptive phenotypes [**154**, **155**], including earlier flowering times for the shorter growing season [**151**], purple anthocyanin-based pigmentation which shields DNA from UV damage [**156**] and increases solar heat absorption [**157**], and macrohairs on the leaf and stem sheath, which are thought to increase herbivore defense [**158**] and/or heat maintenance in colder environments [**159**]. Earlier studies using 50K SNP-chip data for highland populations [**5**] or genomewide data for a small number of individuals [**160**, **161**], have shown that highland maize landraces have experienced significant admixture from *mexicana*, reaching high frequency at some loci, consistent with adaptive introgression.

While some highland and locally-adapted alleles may be beneficial to maize, many introgressed *mexicana* alleles, especially those affecting domestication traits, should be selected against by farmers growing maize landraces. In addition, maize alleles introgressed into *mexicana* should be selected against because maize has accumulated genetic load from reduced population sizes during domestication [**160**] and because domestication traits generally reduce fitness in the wild [**162**,**163**,**164**], e.g. loss of disarticulation and effective seed dispersal [**154**].

In this study, we generate whole genome sequencing to investigate genomic signatures of admixture and selection in paired maize landrace and sympatric *mexicana* populations, sampled from 14 locations across an elevational gradient in Mexico. This expanded sampling of sympatric maize and *mexicana* populations across Mexico, combined with genomewide data and a well-parameterized null model, improves our ability to more formally test for adaptive introgression and identify likely source populations. The source of introgression is of interest, as teosinte demonstrates local adaptation to different niches within the highlands and there is significant genetic structure between *mexicana* ecotypes [**149**,**154**,**165**,**166**,**167**]. Thus we can test whether local *mexicana* populations are the ongoing source for geographically-restricted locally adaptive haplotypes. We use this comprehensive genomic dataset to characterize the bi-directional timing and origin of introgression and evaluate the patterns and scale of natural selection for and against admixture between these taxa.

Results and Discussion

Genomewide *mexicana* ancestry is structured by elevation. We sampled paired sympatric populations from 14 geographically dispersed locations to assess the extent of gene flow between maize and *mexicana* in Mexico. Maize today is grown across the entire elevational range of its wild relatives, from sea-level up to 4000 meters [**168**]. Our sampled sites range from 1547-2600 meters in elevation, which spans a large portion of *mexicana*'s range and exceeds the upper elevational range for maize's wild ancestor, *parviglumis* (Fig 2.1). For each of 14 maize/*mexicana* sympatric sample locations, we resequenced 7-15 individuals per subspecies. We additionally sequenced 43 individuals from 3 *mexicana* reference populations, totalling 348 low-coverage genomes (mean ∼1x). Two of these reference populations are documented to have no adjacent maize agriculture within the past 50 years, while a third higher elevation population (Amecameca) was chosen

because it grows above the elevational range of *parviglumis*, and thus outside of the historical range of maize. For simplicity we refer to these three populations as an 'allopatric' *mexicana* reference panel in the text, in contrast to our sympatric population pairs, but we note that maize has been grown at high density throughout Mexico and gene flow from maize into *mexicana* is possible at Amecameca, and historically at all three locations. We asses gene flow into these *mexicana* reference populations below. For an allopatric maize reference population, we added 55 previously published high-coverage genomes from Palmar Chico [**169**], which sits below the elevational range of *mexicana*.

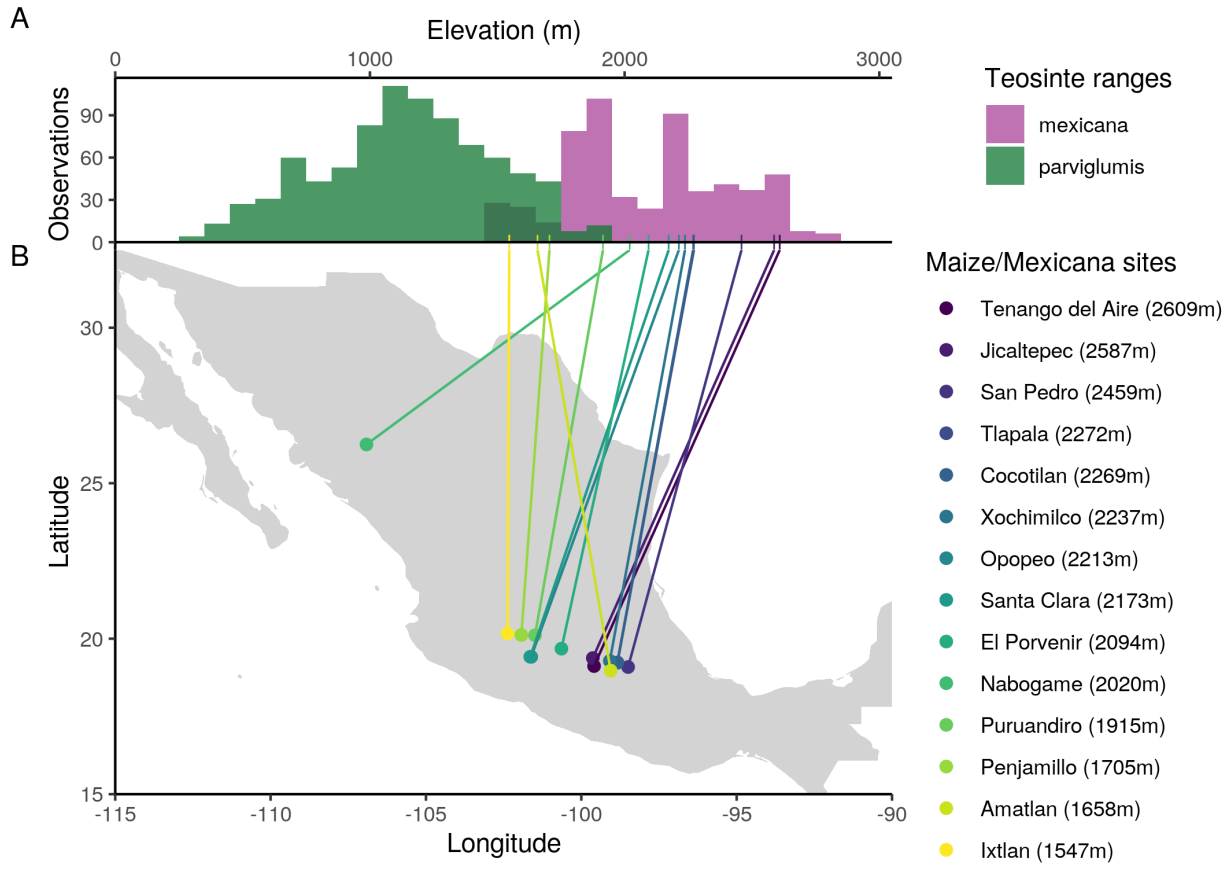


FIGURE 2.1. **Sampled sympatric maize/*mexicana* populations compared to distribution of teosintes** (A) Elevational range of teosintes based on historical occurrence data (1842-2016) from [**146**]. (B) Geographic location and elevation of contemporary sympatric maize and *mexicana* population pairs sampled across 14 sites in Mexico.

Principal components analysis of genetic diversity clearly separates maize and *mexicana*, with putative admixed individuals from sympatric populations having intermediate values along PC1 (PCAngsd, Fig 2.9).

To estimate genomewide ancestry proportions for each individual, we ran NGSAdmix [**52**] with K=2 genetic clusters and genotype likelihoods for all maize and *mexicana* individuals. The two genetic clusters clearly map onto maize and *mexicana* ancestry, with no indication of gene flow into the allopatric maize reference population but small amounts of maize ancestry in two of the three allopatric *mexicana* populations (Fig 2.2A).

Furthermore, we find a positive association between ancestry proportion and elevation (km), with higher *mexicana* ancestry at higher elevations in both sympatric maize ($\beta = 0.196, P = 1.42 \times 10^{-29}$) and sympatric *mexicana* ($\beta = 0.197, P = 4.38 \times 10^{-19}$) individuals (Fig 2.2B).

Increasing *mexicana* ancestry at higher elevations is consistent with selection favoring *mexicana* ancestry at higher elevations, but could also be due to purely demographic processes, e.g. a higher density of (wind-dispersed) *mexicana* pollen at higher elevations, or increased gene flow from non-admixed maize populations at lower elevations. While most populations have admixture proportions well-predicted by their elevation, outlier populations may be the result of recent colonization histories for some locations or adaptation to other environmental niches. Within teosintes, elevation is a major axis of niche separation between *parviglumis* (the ancestor of maize) and *mexicana* [**166**], but genetic differentiation also correlates with soil nutrient content and at least four principal components constructed from climatic variables [**150**].

Origin and timing of introgression. If *mexicana* ancestry found in contemporary landrace genomes facilitated maize's colonization of the highlands approximately 6.2 thousands years ago [**153**], we would expect introgressed ancestry tracts to be short, due to many generations of recombination, and possibly to be derived from an ancient source population common to many present-day maize populations. To test these predictions, we estimated local ancestry across the genome for individuals from each sympatric maize and *mexicana* population using a hidden Markov model (HMM) based on read counts ([**53**] see methods). For each admixed population, this HMM simultaneously estimates local ancestry and, by optimizing the transition rate between different (hidden) ancestry states, the generations since admixture.

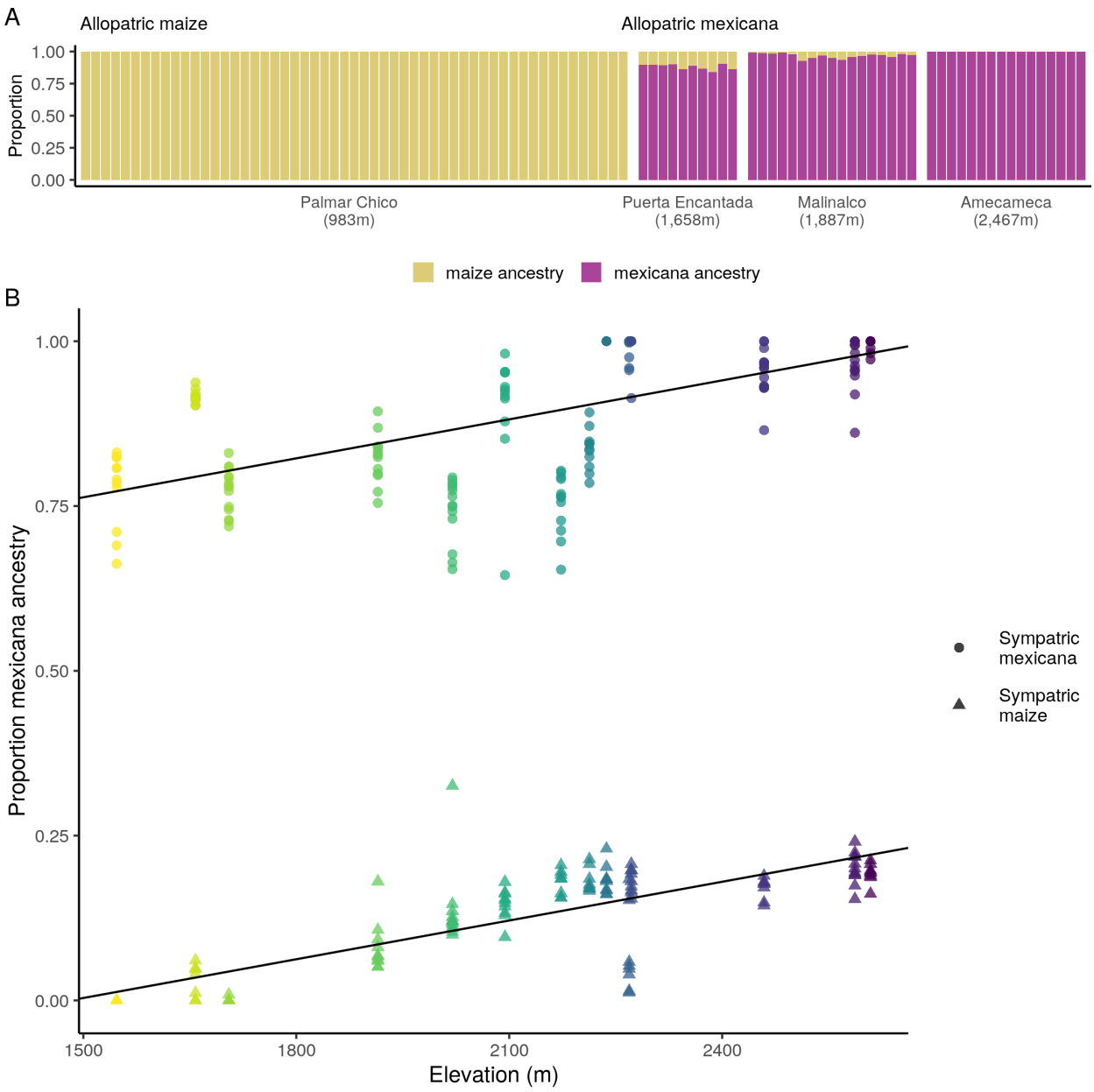


FIGURE 2.2. **Distribution of *mexicana* ancestry by elevation** (A) Genome-wide ancestry estimates (NGSAdmix) for allopatric maize and *mexicana* reference individuals, grouped by sampling location. (B) Genomewide *mexicana* ancestry estimates (NGSAdmix) for sympatric maize and *mexicana* individuals (n = 305) along an elevational gradient, colored by sampling location. Lines show best linear model fit for *mexicana* ancestry by elevation for each subspecies separately.

Admixture is generally old, with median estimates of 1203 generations for sympatric maize populations and 718 generations for sympatric *mexicana* populations (Fig 2.10).

Because this HMM fits a single-pulse model to what was almost certainly multiple admixture events over time, we caution against over-interpretation of exact dates. Multiple pulses or ongoing gene flow biases estimates towards the more recent pulse(s) [**170**, **171**] and even old estimates do not exclude the possibility of limited more recent admixture.

These single-pulse approximations do, however, provide evidence that a large proportion of the introgression, especially into maize, is found on short ancestry tracts and therefore relatively old.

To identify likely source population(s) for introgressed ancestry, we compared F_{ST} between all sympatric populations using only reads from high-confidence homozygous ancestry tracts (posterior > 0.8) for maize and *mexicana* ancestry separately. We find that most *mexicana* ancestry in maize resembles other *mexicana* ancestry introgressed into other maize populations, rather that *mexicana* ancestry from the local sympatric *mexicana* population (Fig 2.3). This finding is consistent with most introgressed ancestry being drawn from a communal source population, but none of the sympatric *mexicana* populations have low enough F_{ST} to tracts introgressed into maize to be a recent source. While we cannot rule out recent introgression from an unsampled source population, the timing of our admixture estimates is more consistent with divergence of *mexicana* ancestry, once introgressed into a maize background, from its original source population(s) (Fig 2.10). Additionally, *mexicana* ancestry tracts in maize have only slightly reduced genetic diversity (π, Fig 2.11), meaning many *mexicana* haplotypes have introgressed into maize at any given locus, with no evidence of a strong historical bottleneck.

Two lower elevation maize populations are an exception to this general pattern: Ixtlan and Penjamillo. These populations have higher F_{ST} between their introgressed ancestry tracts and other *mexicana* tracts in maize (Fig 2.3), more recent timing of admixture estimates (Fig 2.10), and reduced genetic diversity (Figs 2.11-2.12). These patterns could be caused by small population sizes and more recent independent admixture, although F_{ST} does not identify a likely *mexicana* source population. Consistent with this interpretation, we have evidence that local maize at Ixtlan is at least partially descended from recently introduced commercial seed (relayed by local farmers [**5**]).

The lack of a clear reduction in F_{ST} for *mexicana* ancestry tracts between sympatric population pairs, combined with older timing of admixture estimates, indicates that while contemporary

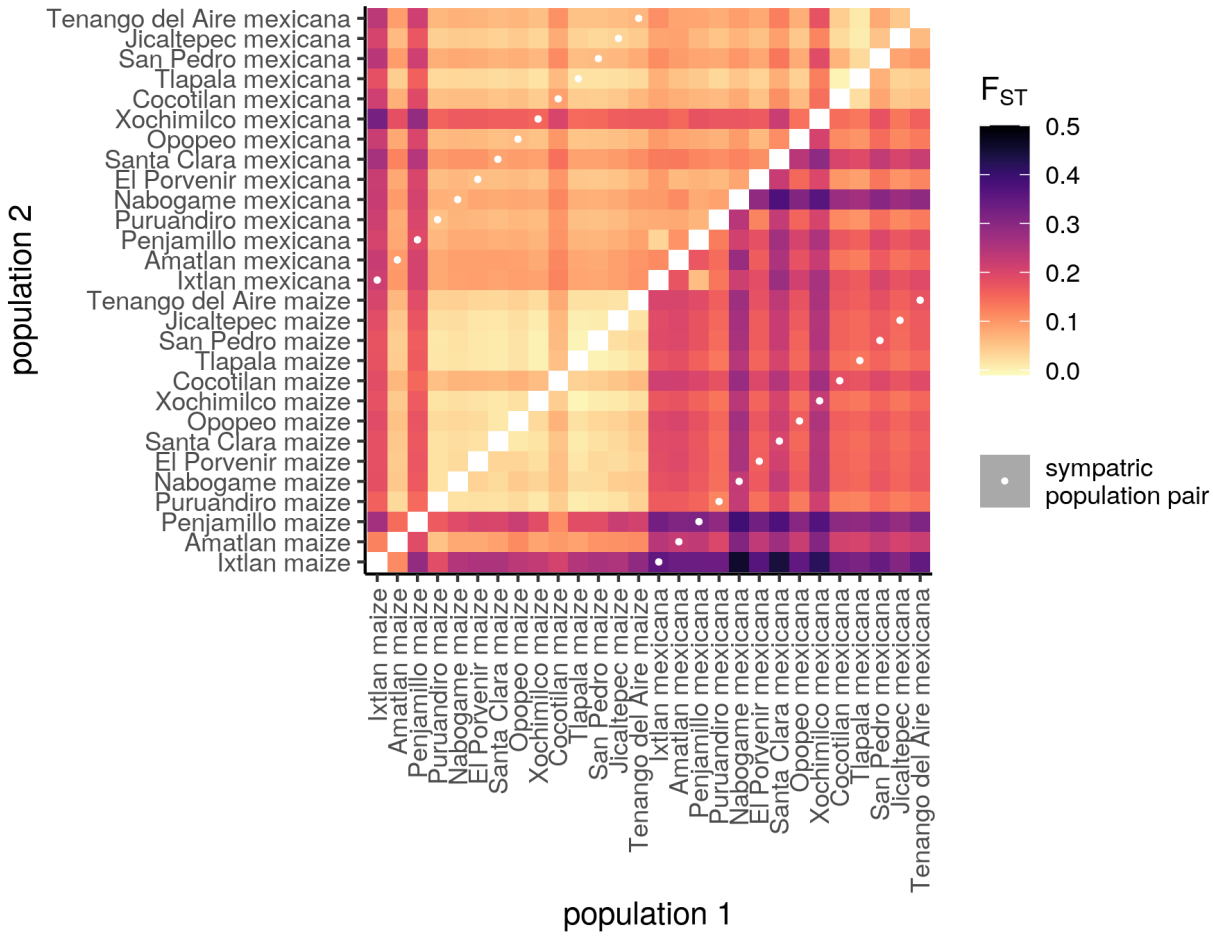


FIGURE 2.3. F_{ST} **between ancestry tracts from different populations** Pairwise F_{ST} between maize ancestry tracts from population 1 (x-axis) and population 2 (y-axis) are shown in the upper left triangle, while F_{ST} estimates for *mexicana* ancestry tracts are shown in the lower right triangle. Populations are sorted by subspecies, then elevation. Local sympatric maize-*mexicana* population pairs are highlighted with a white dot and do not show reduced F_{ST} relative to other (non-local) maize-*mexicana* comparisons. Additionally, introgressed *mexicana* ancestry shows low differentiation between maize populations (creating a light-colored maize block in the lower right triangle) and no potential *mexicana* source populations show especially low F_{ST} with this block. Light coloring generally across the upper left triangle reflects the low differentiation within maize, providing little information to distinguish between potential maize ancestry sources.

hybridization may occur in the field between maize crops and adjacent *mexicana* populations, this is not the source for the bulk of the introgressed ancestry segregating in highland maize.

Instead, we propose that the majority of *mexicana* ancestry in maize derives from admixture over 1000 years ago, possibly from a diverse set of *mexicana* source populations over a large geographic and temporal span, and the resulting ancestry tracts are now distributed across different contemporary maize populations. These genomewide average F_{ST} results, however, do not exclude the possibility that particular regions were introgressed from one or more distinct, possibly local, source populations.

While we also analyzed F_{ST} within high-confidence maize ancestry tracts, we found that maize ancestry is too homogeneous to make inferences about potential admixture source populations of maize into *mexicana* (Figs 2.3, 2.12).

Selection against introgression genomewide. When there is widespread selection against introgressing variants at many loci across the genome, selection will more efficiently remove linked ancestry in regions of the genome with lower recombination rates, which creates a positive relationship between local recombination rate and the proportion of introgressed ancestry [**121**, **122**, **123**, **124**, **125**, **126**, **127**, **128**, **129**, **130**, **172**]. To test whether such negative selection is shaping patterns of introgression genomewide in sympatric maize and *mexicana*, we first divided the genome into quintiles based on the local recombination rates for 1 cM windows. We then ran NGSAdmix on the SNPs within each quintile separately, using K=2 clusters, to estimate maize and *mexicana* ancestry proportions. We used a recombination map from maize [**173**], which is likely to be correlated with other *Zea* subspecies at least at the level of genomic quintiles, but a limitation of this analysis is that we do not have a recombination map for hybrid populations which means that e.g. segregating structural inversions will not necessarily show low recombination rates.

Our results from sympatric maize landraces are consistent with selection against *mexicana* introgression at many loci genomewide, resulting in lower introgressed ancestry in regions of the genome with lower recombination rates (Fig 2.4A). We find a positive Spearman's rank correlation between recombination rate quintile and mean introgressed *mexicana* ancestry proportion ($\rho = 1,\ \mathrm{CI}_{95}[0.85,\ 1.00]$), reflecting the fact that introgression increases monotonically across quintiles. A similar analysis using f_4 statistics replicates this result (see methods, Fig 2.13-2.15 and Table 2.4). The higher elevation maize populations show this pattern most starkly; while all individuals have low *mexicana* ancestry for the lowest recombination rate quintile, some high elevation populations

have individuals with over 40% introgressed ancestry for the highest recombination rate quintile (Fig 2.4B). Using a linear-model fit, we found a significant interaction between recombination rate quintile and the slope of ancestry across elevation in sympatric maize (Table 2.5). This is again consistent with low-recombination rate regions having a stronger effect of linked selection reducing *mexicana* ancestry, with higher elevation maize landraces either experiencing larger amounts of gene flow or retaining more ancestry due to adaptive processes in high recombination regions (Fig 2.16).

Because recombination rate is positively correlated with gene density in *Zea* [**174**], we also tested the Spearman's rank correlation between quintiles defined by coding base pairs per cM and their proportion introgressed *mexicana* ancestry. Again we found evidence supporting pervasive selection against introgression (Fig 2.17, $\rho = -1,\ \mathrm{CI}_{95}[-1,\ -0.85]$).

In contrast, sympatric *mexicana* shows an unexpected negative relationship between recombination rate and introgression from maize, with more *mexicana* ancestry (lower introgression) in the highest recombination rate regions of the genome ($\rho = 1,\ \mathrm{CI}_{95}[0.9,\ 1]$). Correlations with coding bp per cM and based on f_4 statistics corroborate this pattern (see 2.17-2.19 and Table 2.6). While one possible explanation is that introgressing maize ancestry is overall beneficial, not deleterious, a similar pattern could also be produced from a number of different distributions of fitness effects for maize alleles, including for example if most maize alleles are deleterious but some have strong beneficial consequences. While maize ancestry in general is not predicted to provide adaptive benefits in teosinte, invasive *mexicana* in Europe shows selective sweeps for maize ancestry at multiple loci that have contributed to its establishment as a noxious weed [**175**] and we speculate that maize could be a source of alleles adapted to human-modified landscapes.

We repeated these analyses using local ancestry calls as our introgression estimates and found a non-significant Spearman's rank correlation between *mexicana* ancestry and recombination rates for 1 cM windows in sympatric maize (Fig 2.20, $\rho = 0.011,\ \mathrm{CI}_{95}[-0.039,\ 0.062]$) and a negative rank correlation between *mexicana* ancestry and recombination rate in sympatric *mexicana* ($\rho = -0.473,\ \mathrm{CI}_{95}[-0.512,\ -0.432]$). Contrasting results between global and local ancestry methods could be a reflection of true evolutionary differences across different time periods; local ancestry methods capture patterns from more recent gene flow that comes in longer ancestry blocks while

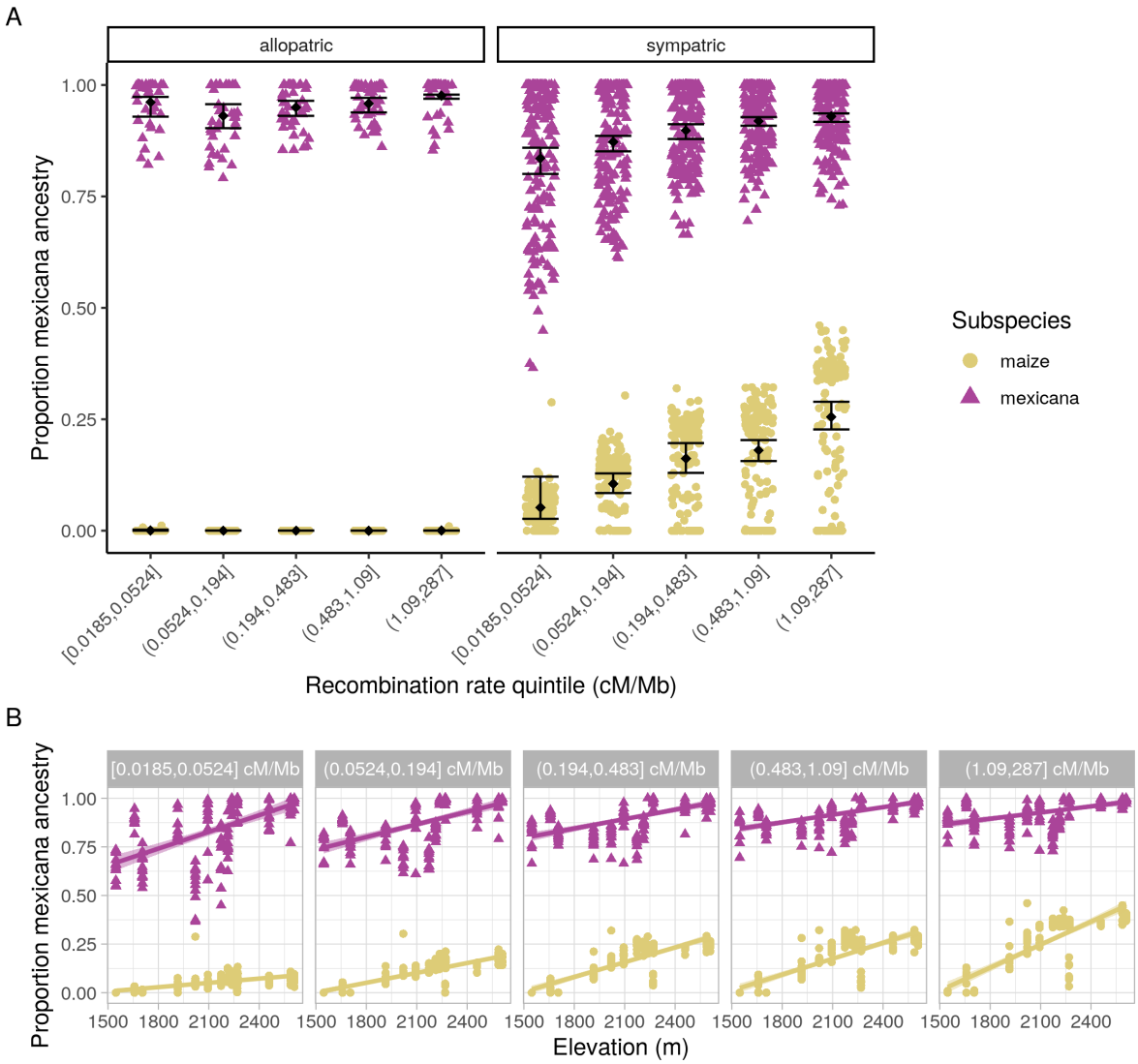


FIGURE 2.4. **(A) *Mexicana* ancestry by recombination rate**. Inferred *mexicana* ancestry in allopatric reference populations (left) and sympatric maize and *mexicana* populations (right) using NGSAdmix (K=2) by recombination rate quintiles. Group mean and 95% confidence interval based on bootstrap percentiles (n = 100) are depicted in black. Ancestry estimates for each individual are shown as points, colored by subspecies, and points are jittered for better visualization. (B) Slopes of *mexicana* ancestry across elevation for each recombination rate quintile, based on NGSAdmix estimates. Each point is a sympatric maize or *mexicana* individual and lines show the best-fit linear model for ancestry by elevation (with shaded 95% confidence interval) estimated separately for each quintile and subspecies.

STRUCTURE-like algorithms (NGSAdmix) and f_4 statistics are based on allele frequencies that collapse information across ancestry blocks of any size, capturing a longer evolutionary time scale.

This interpretation would suggest that *mexicana* has experienced stronger selection against more recent maize gene flow than historical gene flow. However, we caution that local ancestry methods may also have subtle biases in power that are sensitive to local recombination rates and make them less reliable for comparing ancestry patterns across recombination rate quintiles.

Overall, we find support for widespread selection against introgression into maize and mixed results from similar tests of this hypothesis in *mexicana*.

High introgression peaks shared across populations. To assess adaptive introgression in our sympatric populations, we identified introgression 'peaks' where minor ancestry exceeds the genomewide mean by more than 2 standard deviations. We find no strong reduction in average diversity (π) for *mexicana* ancestry at high introgression peaks (Fig 2.11). This maintenance of diversity implies that selection at most peaks has favored multiple *mexicana* haplotypes, and hard sweeps for a recent beneficial mutations on a specific haplotype are rare.

We observe that many high *mexicana* ancestry peaks are shared across subsets of our 14 maize landrace populations (see e.g. chr4, Fig 2.5). While most outlier peaks are unique to a single population, many peaks are shared across 7 or more of the populations (Fig 2.21). To a lesser extent, we also observe sharing of high-introgression peaks for maize ancestry in sympatric *mexicana* populations (Fig 2.22).

High introgression peaks in many independent populations would be very unexpected by chance. However, our sampled populations do not provide independent evidence for adaptive introgression, due to shared gene flow and drift post-admixture (e.g. long-distance human-assisted dispersal of maize seed). To estimate the rate of peak sharing we should expect from demographic processes alone, we simulated 100,000 unlinked loci under a multivariate normal distribution parameterized with the empirical ancestry variance-covariance matrix K (see methods). These simulations preserve the ancestry variance across loci within populations and non-independence in ancestry between populations.

For both sympatric maize and *mexicana*, every population shares an excess of high introgression peaks with all other populations compared to expectations set by our MVN null model. However, peak sharing is most elevated among high elevation maize populations (with the exception of Cocotilan, see Fig 2.6). To investigate the origins of population-specific peaks of introgression, we

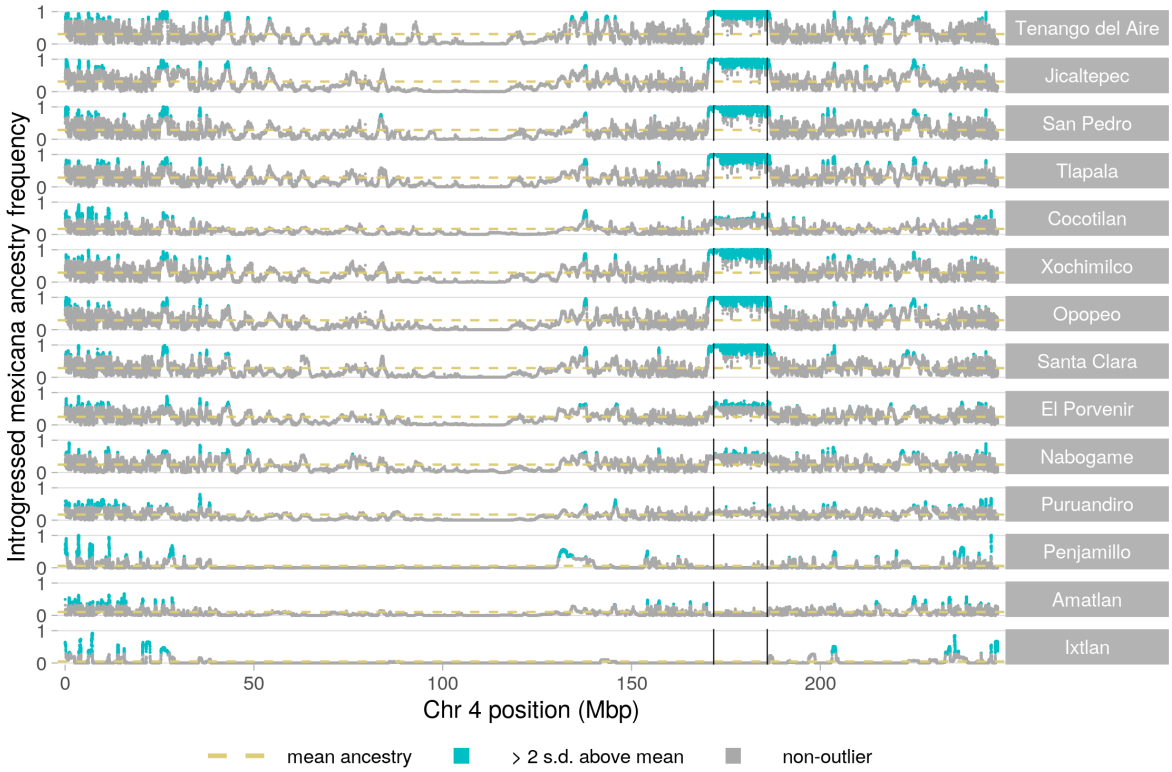


FIGURE 2.5. **Introgression in maize landrace populations across chromosome 4**. Local introgressed ancestry frequency for each maize landrace population compared to their genomewide mean. Populations are ordered from high to low elevation (top to bottom). High introgression peaks with more than 2 standard deviations above the population mean introgressed *mexicana* ancestry are highlighted in blue. Vertical black lines show the previously identified endpoints for a large inversion (*Inv4m* coordinates from Fig 3 of [**166**]). For local ancestry on other chromosomes and for sympatric *mexicana*, see Figs 2.23-2.41

.

calculated F_{ST} between homozygous *mexicana* ancestry in local maize and in each *mexicana* population for these genomic regions. Patterns of F_{ST} between local sympatric pairs at local introgression peaks differed little from background F_{ST} (Fig 2.42), offering little support for the idea that population-specific peaks arose from recent, locally sourced, adaptive introgression. Instead, patterns in maize are consistent with introgressed *mexicana* ancestry tracts from old shared admixture being favored by natural selection, and thus rising to high frequency, in a subset of landraces.

This lack of local adaptive introgression is perhaps surprising given the genetic structure in *mexicana* associated with different ecotypes [**165**] and evidence for local adaptation within teosinte

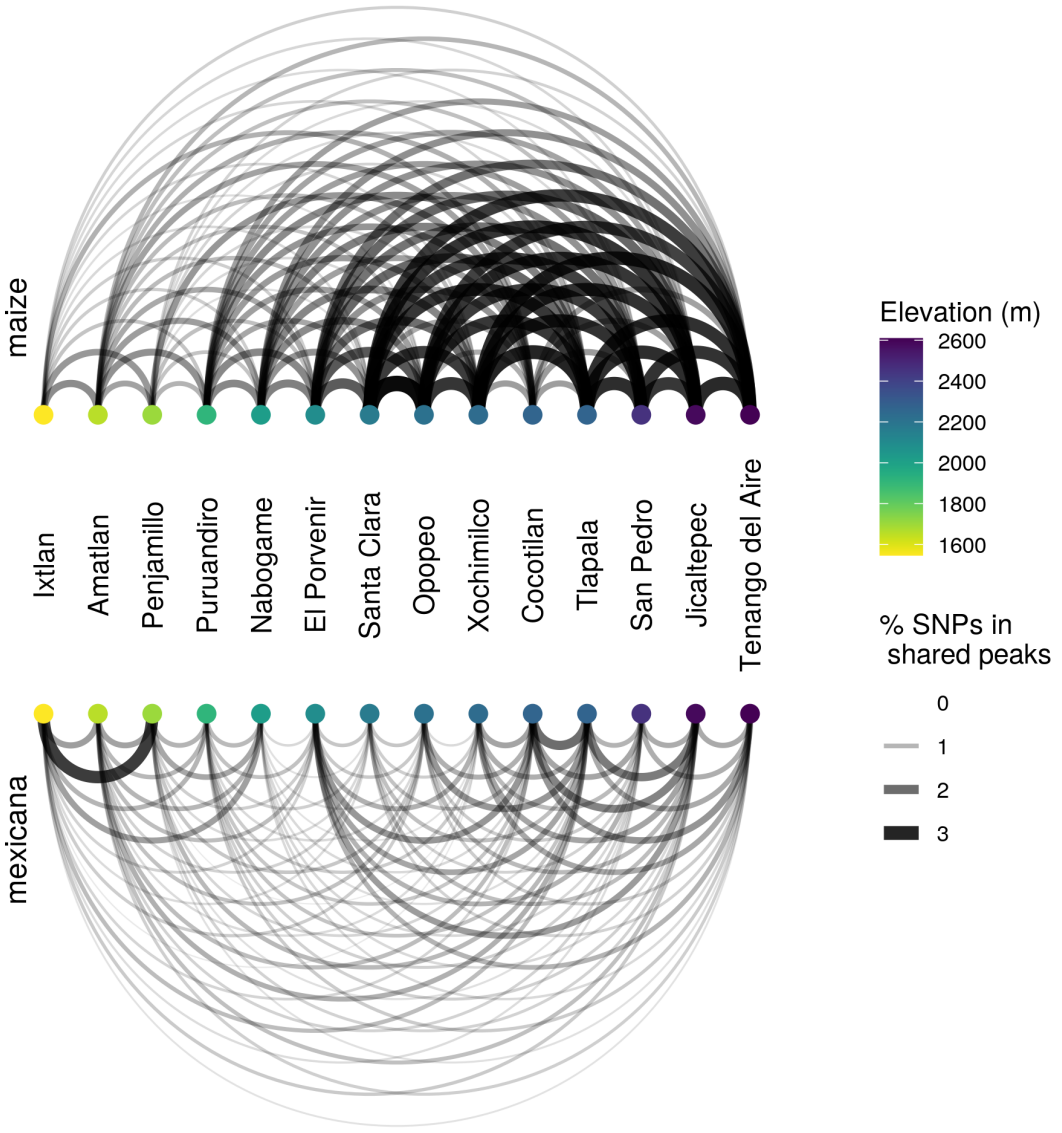


FIGURE 2.6. **Introgression peaks shared across populations** Networks for sympatric maize (top) and *mexicana* (bottom), where each node is a sampled population labelled by location and ordered by elevation. Edges connecting a pair of populations represent the percent of SNPs within shared ancestry peaks (introgressed ancestry > 2 s.d. above each population's mean ancestry). Sharing between all pairs of populations exceeds expectations based on multivariate-normal simulations that model genomewide covariance in ancestry. The relatively darker thicker lines connecting the high elevation maize populations (except for Cocotilan), indicate that these populations share high introgression peaks at especially high frequencies.

across elevation [**148**,**166**,**167**]. However, *mexicana* also has substantial standing variation and we find little evidence for hard sweeps, so one possibility is that local maize and local *mexicana* are adapting to the same environment by different available genetic paths, or even the same causal SNP on a different set of haplotype backgrounds. Older introgressed tracts may also offer more accessible paths for maize adaptation, having already purged some of their linked deleterious variation. Additionally, local exitinction and re-colonization by *mexicana* is common [**154**] and may contribute to a lack of local sourcing of adaptive haplotypes from contemporary *mexicana* populations.

Genomewide scan for selection on introgressed ancestry. We scanned the genome for two types of widespread selection on introgressed ancestry: consistent selection across populations creating an overall excess or deficit of introgression and fitness trade-offs creating steep clines in *mexicana* ancestry across elevation. We used our MVN simulated ancestry frequencies to set false-discovery-rates for excess and deficits of *mexicana* ancestry as well as steeper than expected slopes between *mexicana* ancestry and elevation (see Fig 2.43 for model fit).

We find several regions with high introgression in both directions that are unlikely to be explained by shared demographic history alone (Fig 2.7). These regions of adaptive introgression ($<$ 5% FDR) are spread across the genome and cover a small fraction ($<$0.5%) of the genome in both subspecies. We do not have power to determine if individual genes or regions are barriers to introgression because zero introgressed ancestry is not unusual under our simulated neutral model, given both low genomewide introgression and positive ancestry covariance between admixed populations (Fig 2.7).

Additionally, we identify outlier loci across the genome where *mexicana* ancestry forms steep clines across elevation (Fig 2.7). Our top candidate for strong associations between introgression and elevation in maize is *Inv4m*, a large 14 Mb inversion on chromosome 4 previously identified to have introgressed into high elevation maize landraces [**5**,**160**,**161**,**178**]. This inversion maintains steep elevational clines within teosintes [**166**], overlaps QTLs for leaf pigmentation [**159**] and macrohairs [**159**], and is associated with increased yield in maize at high elevations and decreased yield at low elevations [**178**], but has thus far eluded functional characterization of genes within the inversion [**178**].

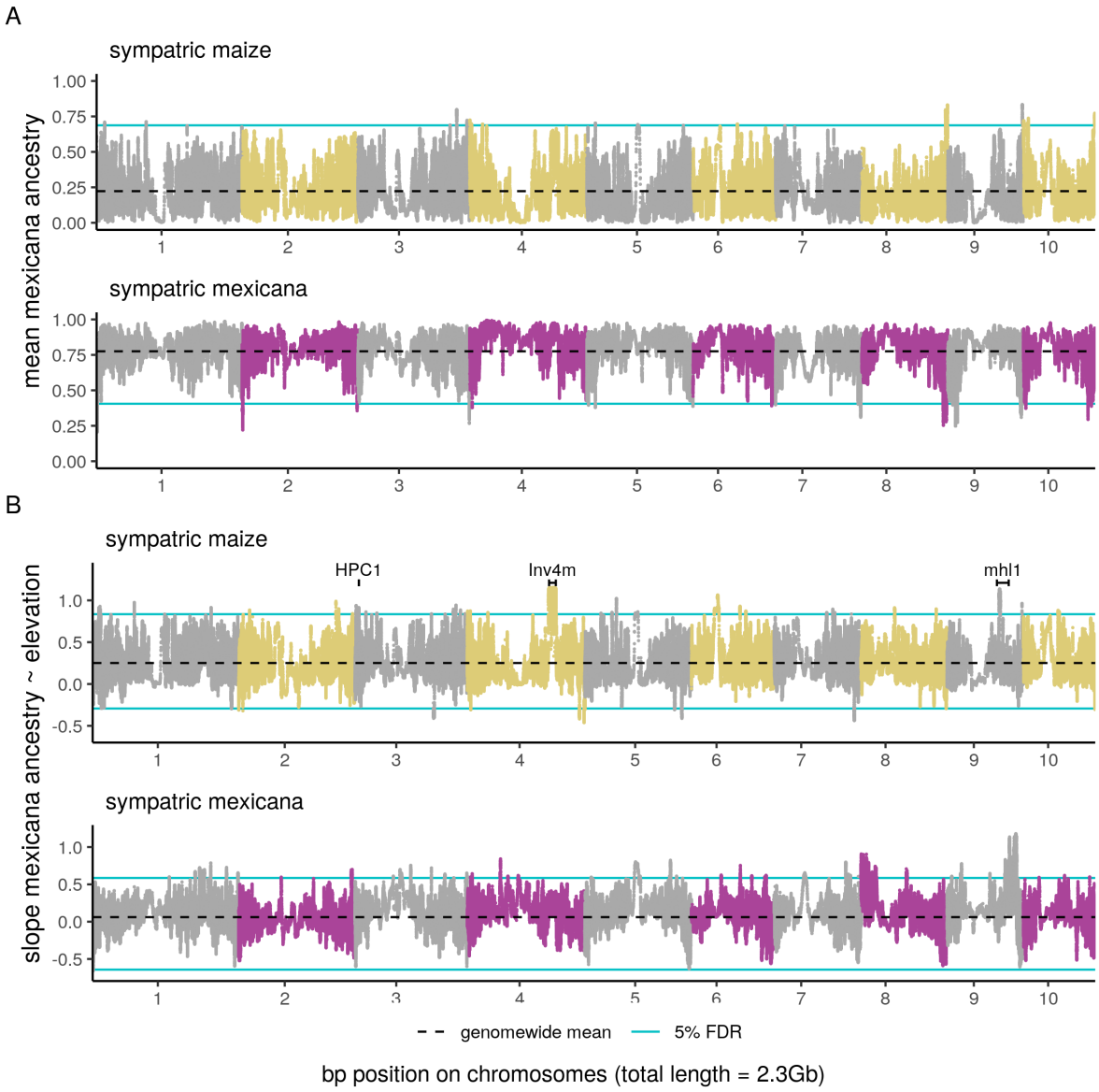


FIGURE 2.7. **Genomewide scan for selection on *mexicana* ancestry**. (A) Mean *mexicana* ancestry in sympatric maize and *mexicana* populations. (B) Slope of *mexicana* ancestry proportion over a 1 km elevation gain in sympatric maize and *mexicana* populations. In both (A) and (B) the blue lines shows the 5% false discovery rates, set using multi-variate normal simulations. Observing *mexicana* ancestry of 0% in sympatric maize or 100% in sympatric *mexicana* was not unexpected based on simulations. Positions for *Inv4m* [**166**] and the *mhl1* locus [**176**] were converted to the maize reference genome v4 coordinates using Assembly Converter (ensembl.gramene.org). Chromosome numbers are placed at the centromere midpoint (approximate centromere positions are from [**177**]).

Our second strongest association co-localizes with *macrohairless1* (*mhl1*), a locus on chromosome 9 that controls macrohair initiation on the leaf blade [**176**] and is associated with a major QTL explaining 52% of macrohair variation between high and low elevation teosinte mapping parents [**159**]. Within teosintes, populations of the lowland ancestor of maize, *parviglumis*, show convergent soft sweeps at the *mhl1* locus not shared by *mexicana* [**149**]. Macrohairs are characteristic highland phenotypes in teosinte and maize and are thought to confer adaptive benefits through insect defence and/or thermal insulation [**158**,**159**]. We identified a 3 Mb outlier region within the larger *mhl1* QTL which we analyzed further using PCA. We found three genetic clusters along the first principal component, evidence that an inversion polymorphism (hereafter *Inv9f*) maintains differentiation between maize/*parviglumis* and *mexicana* haplotypes across this region (Figs 2.44, 2.45). Additionally, principal component two at this inversion separates haplotypes genotyped in maize vs. those genotyped in *mexicana*, consistent with the *mexicana* allele at this inversion introgressing into maize long enough ago to accumulate maize-specific variation, and subsequently sorting in frequency across contemporary maize populations.

The clinal patterns of admixture that we observe at inversions *Inv4m* and *Inv9f* suggest they contribute to elevation-based adaptation in maize, with variation in their fitness impacts even within the historic elevational range of *mexicana*.

While our highest peaks localize with regions previously associated with characteristic highland phenotypes, many additional outlier regions with steep increases in *mexicana* ancestry across elevation have undiscovered associations with local adaptation to elevation. Additionally, outliers for steep ancestry slopes across elevation in sympatric *mexicana* suggest that introgression from maize into *mexicana* may facilitate adaptation in *mexicana* to the lower end of its elevational range.

Selection at candidate domestication genes. We hypothesized that domestication genes will be barriers to introgression bilaterally between maize and *mexicana* [**5**]. While we do not have power to identify individual outlier genes that have low introgression, we can test for enriched overlap between 'introgression deserts' and a set of putative domestication genes spread across the genome.

We examined introgression for a sample of 15 well-characterized domestication genes from the literature (see Table 2.8), and compared them to the regions of the genome with the lowest

5% introgression genomewide across all sympatric maize or *mexicana* populations ('introgression deserts'). A small but enriched subset of these domestication genes overlap with introgression deserts in sympatric maize (7, $P < 0.001$) and likewise in sympatric *mexicana* (7, $P < 0.001$). Among these candidates, we find that *teosinte branched1* (*tb1*), a key transciption factor that regulates branching vs. apical dominance [**179**, **180**], overlaps introgression deserts in both maize and *mexicana*, consistent with *tb1*'s role at the top of the domestication regulatory hierarchy [**181**].

We also find evidence for reduced introgression into both maize and *mexicana* at *teosinte glume architecture1* (*tga1*) [**182**, **183**] and *brittle endosperm2* (*bt2*) [**184**], which are associated with 'naked' edible grains and starch biosynthesis, respectively. Another eight domestication genes [**184**, **185**, **186**, **187**, **188**, **189**, **190**] have low introgression in one direction only.

Among these, *sugary1* (*su1*) in the starch pathway has low maize ancestry in *mexicana* but shows a steep increase in introgressed *mexicana* ancestry with elevation in maize (< 5% FDR), which suggests this gene has pleiotropic effects on non-domestication traits in maize, with fitness trade-offs across elevation. *Sugary1* mutations modify the sweetness, nutrient content and texture of maize kernels (e.g. sweet corn), but also affect seed germination and emergence at cold temperatures [**191**], candidate pleiotropic effects that could be more deleterious at higher elevations.

The remaining four domestication genes do not overlap introgression deserts in either subspecies despite evidence for their role in domestication: *zfl2* (cob rank) [**192**, **193**, **194**], *ba1* (plant architecture) [**195**], *ZmSh1-5.1+ZmSh1-5.2* (seed shattering) [**190**] and *pbf1* (storage protein synthesis) [**196**]. Despite evidence of introgression at many domestication loci, maize landraces retain all of the classic domestication traits, and *mexicana* populations maintain 'wild' forms. Epistasis for domestication traits [**163**] could help explain this discrepancy if compensatory effects from other loci contribute to maintaining domestication traits in admixed highland maize, or key domestication alleles segregate at moderate frequencies within *mexicana* (but do not have the same phenotypic effects in a teosinte background).

Selection within the flowering time pathway. Flowering earlier is adaptive in high-elevation environments where days are cooler and there are fewer total growing degree days in a season. We therefore expect an excess of introgressed *mexicana* ancestry at flowering time genes that may contribute to adaptive early flowering in highland maize. The *mexicana* allele at *High*

PhosphatidylCholine 1 (*HPC1*) has been shown to reduce days to flowering, confers a fitness benefit in maize at higher elevations, and is introgressed in modern flint maize cultivars from Northern Europe and America [**197**]. Here we show that *HPC1* also overlaps an outlier region with one of the steepest increases in *mexicana* ancestry with elevation across sympatric maize populations (+0.91 *mexicana* ancestry proportion/km, FDR $< 5\%$), consistent with a role in adaptive earlier flowering at higher elevations in Mexican landraces. *HPC1* has low introgression from maize into sympatric *mexicana* across the elevational range of this study, suggesting either that the high-elevation *mexicana* allele does not confer fitness tradeoffs in the teosinte background or alternative segregating *mexicana* alleles maintain fitness at lower elevations.

We also tested for selection within the flowering time pathway more broadly using a set of 849 candidate flowering time genes [**198**, **199**]. Only 1/43 genes from the core flowering time pathway (*ZMM5*) [**198**] and 15/806 other candidate flowering time genes [**199**] (+/- 20kb) overlap outlier regions with steep increases in *mexicana* introgression with increasing elevation ($< 5\%$ FDR) in sympatric maize, which matches expected overlap by chance ($\sim$2%, P = 0.76). Thus the steep clinal introgression pattern at *HPC1* in sympatric maize, indicative of strong fitness trade-offs across elevation, is the exception, not the rule, for flowering-time related genes. While for *HPCI* the effect of the *mexicana* allele on reducing flowering time has been confirmed by CRISPR [**197**], a limitation for other less-characterized genes is that we simply assume the *mexicana* allele reduces flowering time. In addition, flowering time is a highly polygenic trait [**200**], which could reduce the strength of selection at individual genes with smaller effect sizes than *HPC1* to below what we can detect using steep ancestry clines. While it is alternatively possible that *mexicana* alleles would show adaptive benefits across the entire range sampled (moderate to high elevation), we find that only 2/849 candidate flowering time genes overlap high mean *mexicana* introgression outliers at a 5% FDR (P = 0.23).

Conclusion

We conclude that the majority of *mexicana* ancestry introgressed into maize over 1000 generations ago and has subsequently been sorted across an elevational gradient, and by selection within individual populations. Differentiation of *mexicana* haplotypes within maize genomewide (F_{ST}) and

at individual introgressed outlier loci (e.g. *Inv9f* at the *mhl1* locus (PCA)) corroborate this timeline. Despite contemporary observations of F1 hybrids in the field [**154**], there is little evidence of significant recent gene flow in either direction between sympatric maize-*mexicana* population pairs. Intrinsic genetic incompatibilities and partial temporal isolation (offset flowering times) clearly create an incomplete barrier to gene flow. However, while hybrids are very challenging to identify and weed out from maize fields at early life stages, farmers can easily distinguish between maize and hybrids when choosing which cobs to plant for the next season. In the other direction, hybrids in most locations are expected to be partially temporally isolated from *mexicana* and hybrid seeds that do not disarticulate are farmer-dependent for successful dispersal and reproduction, although first-generation backcrosses to *mexicana* have been observed [**154**].

Consistent with domestication loci acting as barriers to introgression, in both maize and *mexicana* an enriched subset of candidate domestication genes overlap 'introgression deserts.' More generally, we find introgressed *mexicana* alleles are on average deleterious in maize, but less evidence for a genomewide effect of selection against introgression into *mexicana*, possibly because epistasis masks the impact of maize alleles in a *mexicana* background [**163**]. Some loci show exceptional ancestry patterns consistent with selection favoring introgression in multiple populations, especially for *mexicana* ancestry in the highest elevation maize. While these shared signatures of adaptive introgression are the most striking, the majority of ancestry peaks are exclusive to a single population. Despite this signature of geographically-restricted local adaptation from *mexicana* ancestry, there is no evidence of local population sources for locally adapted haplotypes at these peaks. Thus both broad and local adaptation of maize throughout the highlands appears to have been driven primarily by the sorting of old introgression.

Materials and Methods

Population sampling. We used maize and *mexicana* seed accessions sampled from locations across Mexico in 2008 [**5**] and currently stored at UC Davis. We included 14 maize and 14 *mexicana* accessions that are paired populations sampled in sympatry from the same locations: Ixtlan*, Amatlan, Penjamillo, Puruandiro*, Nabogame*, El Porvenir*, Santa Clara*, Opopeo*, Xochimilco*, Cocotilan, Tlapala, San Pedro*, Jicaltepec and Tenango del Aire* (see Table 2.1). A previous

study of crop-wild admixture genotyped different maize and *mexicana* individuals from 9 of these locations (marked with *), using the Illumina MaizeSNP50 Genotyping BeadChip [**5**]. In addition, we chose three population accessions to sequence as a *mexicana* reference panel: Puerta Encantada and Malinalco were chosen because they have no record of contemporary maize agriculture nearby and a third population, Amecameca, was added as a complement to these two reference populations because it grows at a higher elevation, beyond the historical range of *parviglumis*.

At each sampling location, multiple ears from maternal plants were collected for seed. Population accessions varied in the number of maternal plants with viable seeds. When available, we planted multiple seeds within each ear but only randomly selected one individual for sequencing from the plants that successfully germinated in the greenhouse.

DNA extraction and sequencing. We extracted DNA from leaf tissue and then prepared sequencing DNA libraries using a recently published high-throughput protocol ("Nextera Low Input, Transposase Enabled protocol" [**90**]) with four main steps: (1) DNA shearing and tagmentation by the Nextera TD enzyme, (2) PCR amplification (Kapa2G Robust PCR kit) and individual sample barcoding (custom 9bp P7 indexing primers) (3) library normalization and pooling, and (4) bead-based clean-up and size-selection of pooled libraries. We sequenced the resulting pooled libraries using multiple lanes on Illumina HiSeq 4000 and Novaseq 6000 machines (paired-end 150 bp reads).

To address low sequencing output from some libraries, we re-sequenced 26 libraries (and merged output) and replaced 53 lower-coverage libraries with a higher-coverage library prepared from another seed grown from the same half-sibling family. We excluded 7 samples from analysis because their final libraries did not yield sufficient sequencing output (<0.05x coverage after filtering reads for mapping quality). We additionally removed one lane of sequencing (58 samples) from the study after determining a labelling error had occurred for that plate.

In total, we obtained whole genome sequences for 348 individuals (1.0x average coverage, range: 0.1-2.4x). Of these samples, 43 are *mexicana* from three allopatric populations, with a total of 34.1x combined coverage. The remaining samples are maize and *mexicana* from sympatric populations, 262 of which have sufficient coverage for local ancestry inference ($\geq$ 0.5x, 6-12 per sympatric population, see Fig 2.8). Raw sequencing reads for these low-coverage maize and *mexicana* genomes are available at NCBI (PRJNA657016).

Reference genome and recombination map. We used version 4 of the B73 maize reference genome [**177**] (Zea_mays.B73_RefGen_v4.dna.toplevel.fa.gz, downloaded 12.18.2018 from Gramene).

To find local recombination rates, we converted marker coordinates from a published 0.2 cM genetic map [**173**] to the v4 maize genome using Assembly Converter (ensembl.gramene.org). We removed any markers that mapped to a different chromosome or out of order on the new assembly, and extended the recombination rate estimates for the most distal mapped windows to the ends of each chromosome. From this map, we used *approx()* in R (v3.6.2 [**201**]) to estimate the cM position for any bp position, based on linear interpolation.

Read mapping and filtering. First, we checked read quality using fastQ Screen (v0.14.0 [**202**]) and trimmed out adapter content from raw sequencing reads using the trimmomatic wrapper for snakemake (0.59.1/bio/trimmomatric/pe) [**203**]. We mapped trimmed reads to the maize reference genome using bwa mem (v0.7.17 [**204**]). We then sorted reads using SAMtools (v1.9 [**94**]), removed duplicates using picardtools (v2.7.1) MarkDuplicates and merged libraries of the same individual sequenced on multiple lanes using SAMtools merge. In all subsequent analyses in the methods below we filtered out reads with low mapping scores (< 30) and bases with low base quality scores (< 20).

High-coverage *Tripsacum* genome sequencing. In addition to low-coverage genomes for maize and *mexicana*, we selected a *Tripsacum dactyloides* individual as an outgroup and sequenced it to high coverage. This individual is an outbred 'Pete' cultivar (rootstock acquired from the Tallgrass Prairie Center, Iowa, USA). We extracted genomic DNA from leaf tissue using the E.Z.N.A.® Plant DNA Kit (Omega Biotek), following manufacturer's instructions, and then quantified DNA using Qubit (Life Technologies). We prepared a PCR-free Truseq DNA library and sequenced it with an Illumina HiSeq2500 rapid run (paired-end 250 bp reads). We generated a total of 136.53 Gb of sequencing for this individual, available at NCBI (SRR7758238). For the following analyses that use *Tripsacum* as an outgroup, we randomly subsampled 50% of reads using seqtk, for approximately 30x coverage. We mapped reads to the maize reference using the pipeline described above,

and additionally capped base quality scores with the 'extended BAQ' model in SAMtools [**205**], which reduces the influence of bases with lower alignment quality.

Additional genomes from published sources. For an allopatric maize reference population, we used 55 previously published high-coverage maize genomes from a Tuxpeño landrace grown near Palmar Chico (NCBI: PRJNA616247 [**169**, **206**]). This maize population grows at 983 m, below the elevational range for *mexicana*.

For a parviglumis reference population, we used 50 previously published high-coverage lowland individuals sampled from the 'Mound' population at 1,008 m near Palmar Chico [**169**, **206**, **207**] (NCBI: PRJNA616247, see Table 2.2). We mapped and filtered reads for these individuals using the pipeline described above and capped base quality scores using BAQ.

SNP calling. We called SNPs using a combined panel of the 348 low-coverage genomes sequenced in this study for sympatric maize, sympatric *mexicana*, and allopatric *mexicana*, and the 55 high-coverage allopatric maize reference genomes described above. We used ANGSD (v0.932 [**95**]) to identify variant sites with minor allele frequencies $\geq 5\%$ in the total sample based on read counts ('angsd -doMajorMinor 2 -minMaf 0.05 -doCounts 1 -doMaf 8'). In addition to mapping and base quality filters ('-minMapQ 30 -minQ 20'), we capped base qualities using the extended per-Base Alignment Quality algorithm ('-baq 2' [**205**]) and removed sites that did not have at least 150 individuals with data or had sequencing depth exceeding 2.5x the total sample mean depth. To apply this total depth filter, we estimated mean depth ('angsd -doCounts 1 -doDepth 1 -maxDepth 10000') for 1000 regions of length 100bp randomly sampled using bedtools (v2.29.0 [**108**]). In total, we identified 52,118,357 SNPs on the assembled chromosomes. In conjunction with SNP calling, we produced genotype likelihoods for each individual at these variant sites using the SAMtools GL method [**94**] implemented in ANGSD ('-GL 1 -doGlf 2').

Global ancestry inference. To estimate genetic relationships between populations and their genomewide ancestry proportions, we used methods specific to low-coverage data that rely on genotype likelihoods, rather than called genotypes. Because these methods are sensitive to SNPs in high linkage disequilibrium (LD), we thinned genotype likelihoods to every 100th SNP ($\sim$4kb spacing) [**208**]. To confirm that maize and *mexicana* subspecies form a major axis of genetic

variation in our sample, we estimated the genetic covariance matrix between individuals using PCAngsd (v0.98.2 [**97**]) and visualized principal components computed using *eigen()* in R. We then estimated global ancestry proportions using the same thinned genotype likelihood files as input to NGSAdmix [**52**], using K = 2 clusters. Clusters clearly mapped onto the two reference groups, which we used to label the two ancestry components as 'maize' and '*mexicana*'.

Local ancestry and timing of admixture. We inferred local ancestry across the genome using a hidden Markov model that is appropriate for low-coverage data because it models genotype uncertainty down to the level of read counts for all admixed individuals (ancestry_hmm [**53**]). This method relies on allele counts from separate reference populations to estimate allele frequencies for each ancestry. Because some of our reference individuals have too low of coverage to accurately call genotypes, we randomly sampled one read per individual to get unbiased frequency estimates for major and minor alleles at each site ('angsd -doCounts 1 -dumpCounts 3'). To maximize ancestry-informativeness of sites in this analysis, we identified SNPs with allele frequency differences of at least 0.3 between subspecies ('angsd -doMajorMinor 3 -GL 1 -baq 2 -doMaf 1') estimated from at least 44 reference maize and 12 reference *mexicana* individuals with sequencing coverage at a site. We then calculated genetic distances between SNPs using the maize recombination map and filtered our enriched variants to have minimum 0.001 cM spacing between adjacent SNPs .

Running ancestry_hmm jointly infers local ancestry for each individual and the time since admixture. This HMM method assumes a neutral demographic history in which a constant-size admixed population was formed by a single admixture event t generations in the past, and finds the t that maximize the likelihood of the observed read counts and hidden local ancestry state across each admixed individual's genome. The timing of admixture defines the generations for possible meiotic recombination between ancestry tracts, and therefore scales the transition probabilities between hidden ancestry states. In addition to t, the HMM outputs the posterior probabilities for homozygous maize, homozygous *mexicana*, and heterozygous ancestry for each individual at every site. We analysed each sympatric maize and *mexicana* population separately, using the population's mean NGSAdmix global ancestry estimate as a prior for mixing proportions, 100 generations as a prior for admixture time (range: 0-10000), an approximate effective population size (Ne) of 10,000 individuals, genetic positions for each SNP based on the maize linkage map, and an estimated

sequencing base error rate of 3×10^{-3}. We ran ancestry_hmm with an optional setting to bootstrap 100 random samples of 1,000-SNP genomic blocks to estimate uncertainty around the estimated generations since admixture (t). To test the sensitivity of the HMM to our choice of Ne, we re-ran ancestry_hmm with two other Ne's that differ by an order of magnitude (Ne = 1k, 100k), but did not analyze these results further after finding high correspondence for both local ancestry and timing estimates.

To get a single point estimate for local ancestry at a site for an individual, we computed a sum of *mexicana* ancestry from the different possible ancestry states, weighted by their posterior probabilities: *mexicana* ancestry proportion = P(homozygous *mexicana*) + $^1/_2$P(heterozygous maize-*mexicana*). In addition, for analyses that require ancestry tract positions, we assumed that the estimated ancestry at a focal site extends halfway to the next site with a local ancestry estimate.

Diversity within ancestry. Using *mexicana* ancestry estimates from the HMM, we identified high-confidence homozygous ancestry tracts for both maize and *mexicana* ancestry (posterior $>$ 0.8). We filtered individual bams for reads that overlap these tracks and used the resulting filtered bams to calculate diversity within both maize and *mexicana* ancestry, separately. We estimate diversity using the ANGSD/realSFS framework which is appropriate for low-coverage sequence data it takes into account uncertainty in both genotypes and variant sites. We created a concensus fasta sequence for *Tripsacum* ('angsd -doFasta 2') to use as the ancestral state for polarizing the unfolded site frequency spectrum in these analyses.

For each population and ancestry, we estimated the site allele frequencies ('angsd -doSaf 1 -GL 1') and subsequently estimated the genomewide site frequency spectrum (SFS). We then used this SFS as a prior to estimate within-ancestry pairwise diversity (π) genomewide from the site allele frequencies ('realSFS saf2theta').

For each pair of populations and ancestry, we additionally used realSFS to estimate the two dimensional SFS from the individual population site allele frequencies genomewide. We then used this 2D SFS as a prior to estimate genomewide within-ancestry F_{ST} between the two populations ('realSFS fst index -whichFst 1'). This call uses Hudson's F_{ST} estimator [**209**] as parameterized in [**210**].

Effect of local recombination rate on introgressed ancestry. To estimate the effects of linked selection and recombination rate on genomewide introgression patterns, we compared *mexicana* ancestry estimates across genomic quintiles. Based on a 0.2 cM-resolution recombination map [**173**] for maize, we merged adjacent recombination windows into larger 1 cM non-overlapping windows and calculated each window's mean recombination rate and overlap with coding base pairs (bedr 'coverage') [**211**]. We retrieved coding regions ('CDS') using gene annotations from Ensembl (ensemblgenomes.org, Zea_mays.B73_RefGen_v4.41.chr.gff3.gz, dowloaded 11.6.2018). We sorted windows into quintiles for either recombination rate or coding density (bp/cM). Each quintile covers approximately 1/5 of the genome based on physical bp.

i. NGSAdmix estimates. To estimate ancestry proportions for each recombination rate quintile, we first reduced LD by thinning to 1% of SNPs (every 100th) and ran NGSAdmix 5 times separately (once per quintile) with K=2 clusters. We assigned 'maize' and '*mexicana*' labels to the ancestry clusters based on majority assignment to the respective allopatric reference panels. To bootstrap for uncertainty, we re-sampled 1 cM windows with replacement from each quintile 100 times, and re-ran NGSAdmix on the resulting bootstrap SNP sets. Using these results, we calculated 95% percentile bootstrap confidence intervals for the estimated admixture proportions, and the Spearman's rank correlation between the recombination rate (or coding bp per cM) and admixture proportion ranks for each quintile. We also tested for a difference in ancestry slopes with elevation across different recombination rate quintiles by fitting a linear model with an elevation by recombination quintile interaction term: *mexicana* ancestry $\sim$ elevation + r + elevation*r. Using lm() in R, we fit this model for sympatric maize and sympatric *mexicana* separately, treating quintiles as a numeric scale 0-4.

f_4 estimates. In a complementary analysis, we used a ratio of f_4 statistics as an alternative method to estimate ancestry proportions by quintile. The f_4 statistic measures shared genetic drift (allelic covariance) between populations in a phylogeny, due to either shared branch lengths or admixture events in the evolutionary history relating these populations. Excess shared drift with one population from a pair of sister populations in the tree is a signature of admixture, analogous

to the ABBA-BABA test [**212**], and a ratio of two f_4 statistics can be used to quantify the admixture proportion. Assuming the basic phylogenetic tree (((*parviglumis*, allopatric maize), allopatric *mexicana*), *Tripsacum*) in Fig 2.13, we can estimate α, the proportion of ancestry inherited from *mexicana* in an admixed population, as follows [**212**, **213**]:

$$\alpha = \frac{f_4(\text{Tripsacum, } \textit{parviglumis}\text{; X, allopatric maize})}{f_4(\text{Tripsacum, } \textit{parviglumis}\text{; allopatric } \textit{mexicana}\text{, allopatric maize})}.$$

The denominator of this statistic estimates the branch length leading to *parviglumis* and allopatric maize that separates these sister subspecies from allopatric *mexicana*; the full ratio estimates the proportion of this branch that separates sympatric population X from *parviglumis* and allopatric maize, i.e. the *mexicana* ancestry in X. Because the f_4 statistic is sensitive to additional unmodeled admixture within the tree, we limited our allopatric *mexicana* group to individuals from just one of the three reference populations (Amecameca), which showed no evidence of admixture in our global ancestry analysis (see Fig 2.2).

For each 1 cM window across the genome, we used ANGSD to calculate ABBA-BABA statistics from observed read counts for the 4 populations in the numerator and denominator of the α estimator separately ('angsd -doabbababa2 1 -remove_bads 1 -minMapQ 30 -minQ 20 -doCounts 1 -doDepth 1 -maxDepth 10000 -useLast 1 -blockSize 5000000'). From the resulting output files, we summed the negative of the ABBA-BABA numerator ('Num') and divided by the total number of included sites ('nSites') across all 1 cM windows within a quintile to get the f_4 statistic [**214**].

We then calculated the Spearman's rank correlation between the recombination rate quintiles and admixture proportion ranks for these quintiles. We calculated simple bootstrap confidence intervals for our ancestry estimates and correlations by re-sampling 1 cM windows within quintiles with replacement 10,000 times and re-calculating the f_4 ratios and resulting rank correlation across quintiles to construct 95% percentile confidence intervals. We repeated this analysis using quintiles based on coding bp per cM in place of recombination rate (cM/Mbp).

ii. Local ancestry estimates. We also calculated the Spearman's rank correlation between local recombination rate (or coding bp per cM) and local ancestry proportion at the level of individual 1 cM windows. For each window, we averaged local ancestry estimates from the HMM across all individuals within sympatric maize, and separately, sympatric *mexicana*. We then calculated

simple bootstrap confidence intervals for our local ancestry estimates and local recombination rate (or coding bp per cM) by re-sampling 1 cM windows across the genome with replacement 10,000 times and re-calculating the rank correlation across windows to construct 95% percentile confidence intervals.

Local ancestry simulations. We simulated *mexicana* ancestry population frequencies using a multivariate-normal null model:

$$\textit{mexicana} \text{ ancestry} \sim \text{MVN}(\vec{\alpha}, K)$$

where $\vec{\alpha}$ is the vector of mean *mexicana* ancestry frequencies genomewide for each sympatric population and K is the empirical ancestry variance-covariance matrix relating these 14 populations. The diagonal entries of the K matrix capture the expected variation in local ancestry across the genome within populations due to drift and random sampling. The off-diagonals capture ancestry covariances between populations created by shared gene flow and drift post-admixture: at loci where one population has an excess of introgression, other admixed populations with shared demographic history will also tend to have an excess of introgression.

To construct K, we calculated the covariance in ancestry between each pair of populations i and j using all L loci with local ancestry calls genomewide:

$$K[i,j] = \frac{1}{L}\sum_{l=1}^{L}(Anc_{i,l} - \alpha_i)(Anc_{j,l} - \alpha_j).$$

Above, $Anc_{i,l}$ and $Anc_{j,l}$ are local ancestry frequencies at a locus l while α_i and α_j are the mean local ancestry frequencies across the genome for populations i and j.

For sympatric maize and sympatric *mexicana* separately, we calculated the empirical K matrix between populations from all 14 sympatric locations, and then took 100,000 independent draws from their MVN distribution, thereby simulating *mexicana* ancestry for all populations at 100,000 unlinked loci. Because ancestry frequencies are bounded at [0,1] but normal distributions are not, we truncated any simulated values outside of this range.

Introgression peaks shared between populations. To characterize introgression peak sharing between individual populations, we defined 'ancestry peaks' as sites where a population has over 2 standard deviations more introgressed ancestry than the genomewide mean. We counted the number of peaks that are shared between all pairs and combinations of populations. To compare these results to our null model, we also counted the number of introgression peaks shared by populations in our simulated dataset, using the 2 s.d. cutoff set by the empirical data to define peaks.

Because *mexicana* ancestry shows significant diversity, we additionally characterized diversity for *mexicana* ancestry peaks introgressed into maize. For all introgressed ancestry outlier regions in a focal maize population, we used ANGSD to estimate pairwise diversity within the population (π) and differentiation (F_{ST}) between the focal sympatric maize populations and their local sympatric *mexicana* population. We focused on the *mexicana* ancestry within peaks by limiting our diversity estimates to only include high-confidence homozygous *mexicana* ancestry tracts (posterior > 0.8). For these analyses, we pooled information across outlier peaks, but distinguish between introgression peaks exclusive to the focal population and introgression peaks shared between the focal population and at least 3 other sympatric maize populations. We used global estimates of the SFS and 2D SFS as priors to estimate π and F_{ST} for the subsets of the genome within introgression peaks, and otherwise followed the same methods listed above in 'Diversity within ancestry'.

Genomewide scan for ancestry outliers. For sympatric maize and *mexicana* separately, we calculated the mean *mexicana* ancestry across all individuals at a locus, and fit a linear model using lm() in R to estimate the slope of *mexicana* ancestry frequencies for sympatric populations across elevation (km): *mexicana* ancestry $\sim$ elevation. We then repeated these summary statistics for every locus with an ancestry call in the empirical data and each simulated locus in the MVN simulated data.

We calculated 5% false-discovery-rate (FDR) cutoffs for high and low *mexicana* ancestry using the Benjamini-Hochberg method [**110**] and simulation results to estimate the expected frequency of false-positives under our null MVN model (one-tailed tests). We repeated this approach to identify outlier loci with steep positive (or negative) slopes for *mexicana* ancestry across elevation at a 5% FDR.

Test for reduced introgression at domestication genes. To test whether domestication genes are unusually resistant to introgression, we first defined 'introgression deserts' as regions with the lowest 5% of introgression genomewide across all sympatric maize (or, separately, sympatric *mexicana*) populations. We then looked up v4 coordinates on Ensembl.org for genes associated with maize domestication in the literature (Table 2.8), and used bedtools 'intersect' to identify which of these genes ±20 kb overlap introgression deserts. To test for significance, we randomly shuffled the gene positions across the genome (bedtools 'shuffle') 1000 times and re-calculated overlap with introgression deserts for each permuted data set.

Test for selection within the flowering time pathway. We identified a list of 48 core flowering time pathway genes from the literature [**198**], and a broader list of 905 flowering time candidate genes [**198**, **199**]. From the combined set, we included 849 total genes (43 core pathway) which we were able to localize on assembled autosomes of the v4 reference genome using MaizeGDB gene cross-reference files [**215**]. We counted the number of genes ±20 kb that intersected with outlier regions for steep increases in *mexicana* introgression with elevation (and, separately, high *mexicana* introgression) in sympatric maize populations (< 5% FDR) using bedtools 'intersect', then tested for significance by repeating this analysis with 1000 randomly shuffled gene positions.

Analysis pipeline and data visualization. We constructed and ran bioinformatics pipelines using snakemake (v.5.17.0 [**216**]) within a python conda environment (v3.6). We analyzed and visualized data in R (v3.6.2 [**201**]) using the following major packages: tidyverse (v1.3.0 [**58**]), viridis (v0.5.1 [**217**]), bedr (v1.0.7 [**211**]), boot (v.1.3_25 [**218**, **219**]), gridExtra (v2.3 [**220**]), ggupset (v0.3.0 [**221**]) and tidygraph (1.2.0 [**222**]). All scripts can be found on our gitHub repository, https://github.com/ecalfee/hilo, which also includes a full list of software and versions (see envs/environment.yaml).

Acknowledgments

The authors would like to acknowledge funding from NSF award number 1546719 to JRI and GC. This work was also supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH R01 GM108779 and R35 GM136290, awarded to GC). We thank Pesach Lubinsky for collecting the seeds sequenced in this study. We also want to thank the Coop

and Ross-Ibarra labs, and the HILO and Zeavolution working groups for helpful feedback on this work.

Associated Publication

An earlier version of Chapter 2 was posted as a preprint to BioRxiv on March 5, 2021. This preprint will be updated with a link to the final peer-reviewed article upon publication:

Calfee E, Gates D, Lorant A, Perkins TA, Coop G, Ross-Ibarra J. Selective sorting of ancestral introgression in maize and teosinte along an elevational cline. BioRxiv. 2021; doi:10.1101/2021.03.05.434040

Supporting Information

Table 2.1. **Population metadata.**

Subspecies	Group	Location	Country	Elev. (m)	Latitude	Longitude	Accession
mexicana	allopatric	Puerta Encantada	Mexico	1658	18.9725	-99.0298	RIMME0033
mexicana	allopatric	Malinalco	Mexico	1887	18.9531	-99.503	RIMME0020
mexicana	allopatric	Amecameca	Mexico	2467	19.139	-98.7733	RIMME0022
maize	sympatric	Ixtlan	Mexico	1547	20.1683	-102.373	RIMMA0371
maize	sympatric	Amatlan	Mexico	1658	18.9719	-99.0551	RIMMA0369
maize	sympatric	Penjamillo	Mexico	1705	20.1176	-101.93	RIMMA0361
maize	sympatric	Puruandiro	Mexico	1915	20.1076	-101.49	RIMMA0370
maize	sympatric	Nabogame	Mexico	2020	26.2465	-106.915	RIMMA0360
maize	sympatric	El Porvenir	Mexico	2094	19.6789	-100.64	RIMMA0363
maize	sympatric	Santa Clara	Mexico	2173	19.4184	-101.642	RIMMA0362
maize	sympatric	Opopeo	Mexico	2213	19.4181	-101.613	RIMMA0368
maize	sympatric	Xochimilco	Mexico	2237	19.2861	-99.0827	RIMMA0374
maize	sympatric	Cocotilan	Mexico	2269	19.2244	-98.8427	RIMMA0367
maize	sympatric	Tlapala	Mexico	2272	19.2351	-98.8368	RIMMA0365
maize	sympatric	San Pedro	Mexico	2459	19.0886	-98.4935	RIMMA0372
maize	sympatric	Jicaltepec	Mexico	2587	19.3764	-99.6303	RIMMA0366
maize	sympatric	Tenango del Aire	Mexico	2609	19.1197	-99.5896	RIMMA0373
mexicana	sympatric	Ixtlan	Mexico	1547	20.1683	-102.373	RIMME0029
mexicana	sympatric	Amatlan	Mexico	1658	18.9719	-99.0551	RIMME0027
mexicana	sympatric	Penjamillo	Mexico	1705	20.1176	-101.93	RIMME0019
mexicana	sympatric	Puruandiro	Mexico	1915	20.1076	-101.49	RIMME0028
mexicana	sympatric	Nabogame	Mexico	2020	26.2465	-106.915	RIMME0018
mexicana	sympatric	El Porvenir	Mexico	2094	19.6789	-100.64	RIMME0021
mexicana	sympatric	Santa Clara	Mexico	2173	19.4184	-101.642	RIMME0034
mexicana	sympatric	Opopeo	Mexico	2213	19.4181	-101.613	RIMME0026
mexicana	sympatric	Xochimilco	Mexico	2237	19.2861	-99.0827	RIMME0035
mexicana	sympatric	Cocotilan	Mexico	2269	19.2244	-98.8427	RIMME0025
mexicana	sympatric	Tlapala	Mexico	2272	19.2351	-98.8368	RIMME0023
mexicana	sympatric	San Pedro	Mexico	2459	19.0886	-98.4935	RIMME0030
mexicana	sympatric	Jicaltepec	Mexico	2587	19.3764	-99.6303	RIMME0024
mexicana	sympatric	Tenango del Aire	Mexico	2609	19.1197	-99.5896	RIMME0031

Table 2.2. **Parviglumis SRA IDs.**

Run	Isolate	Run (cont.)	Isolate (cont.)
SRR11448802	PC_M59_ID1	SRR13207117	PC_J01_ID1
SRR11448838	PC_I05_ID1	SRR11448791	PC_N48_ID1
SRR11448793	PC_N13_ID1	SRR11448812	PC_L08_ID1
SRR11448797	PC_N09_ID1	SRR11448834	PC_I50_ID1
SRR11448799	PC_N07_ID1	SRR13207120	PC_O08_ID1
SRR11448800	PC_N04_ID1	SRR11448794	PC_N11_ID1
SRR11448803	PC_M58_ID1	SRR11448804	PC_M15_ID1
SRR11448805	PC_M05_ID1	SRR11448807	PC_L56_ID1
SRR11448809	PC_L14_ID1	SRR13207095	PC_J08_ID1
SRR11448811	PC_L12_ID1	SRR13207116	PC_O59_ID1
SRR11448814	PC_K60_ID1	SRR13207160	PC_J14_ID1
SRR11448816	PC_K54_ID1	SRR13207149	PC_J48_ID1
SRR11448818	PC_K02_ID1	SRR13207138	PC_K55_ID1
SRR11448819	PC_J51_ID1	SRR13207131	PC_L06_ID1
SRR11448820	PC_J50_ID1	SRR13207130	PC_L10_ID1
SRR11448823	PC_J13_ID1		
SRR11448824	PC_J12_ID1		
SRR11448825	PC_J10_ID1		
SRR11448827	PC_J04_ID1		
SRR11448830	PC_I58_ID1		
SRR11448832	PC_I52_ID1		
SRR11448836	PC_I08_ID1		
SRR11448837	PC_I06_ID1		
SRR13207106	PC_J07_ID1		
SRR13207118	PC_O51_ID1		
SRR13207119	PC_O10_ID1		
SRR13207121	PC_N60_ID1		
SRR13207122	PC_N58_ID1		
SRR13207123	PC_N57_ID1		
SRR13207124	PC_N56_ID1		
SRR13207125	PC_N14_ID1		
SRR13207126	PC_N10_ID1		
SRR13207127	PC_L48_ID1		
SRR13207128	PC_I53_ID1		
SRR13207129	PC_I11_ID1		

Table 2.3. **Spearman's rank correlation between *mexicana* ancestry (NGSAdmix) and recombination rate (or coding bp per cM) quintiles**

group	feature	Spearman's ρ	2.5%	97.5%
allopatric maize	recombination rate (cM/Mb)	-0.36	-0.90	0.67
allopatric *mexicana*	recombination rate (cM/Mb)	0.40	0.10	1.00
sympatric maize	recombination rate (cM/Mb)	1.00	0.85	1.00
sympatric *mexicana*	recombination rate (cM/Mb)	1.00	0.90	1.00
allopatric maize	coding bp per cM	0.62	-0.30	1.00
allopatric *mexicana*	coding bp per cM	-0.40	-0.90	-0.20
sympatric maize	coding bp per cM	-1.00	-1.00	-0.85
sympatric *mexicana*	coding bp per cM	-0.90	-1.00	-0.80

Table 2.4. **Spearman's rank correlation between ancestry (f_4 ratio) and recombination rate (or coding bp per cM) quintiles in sympatric maize**

group	feature	Spearman's ρ	2.5%	97.5%
sympatric maize	recombination rate (cM/Mb)	1.00	0.30	1.00
sympatric maize	coding bp per cM	-0.90	-1.00	0.10

Table 2.5. **Ancestry by elevation and recombination rate quintile.** Best-fitting linear models for ancestry proportion predicted by an elevation by recombination rate interaction: *mexicana* ancestry $\sim$ elevation + r + elevation*r. Here, r is the recombination rate quintile, treated as numeric [0-4]. This model only uses ancestry estimates for sympatric individuals and is fit separately for maize and *mexicana* samples.

group	term	estimate	std.error	statistic	p.value
sympatric maize	intercept	-0.122	0.027	-4.512	7.67E-06
sympatric maize	elevation (km)	0.083	0.013	6.596	9.01E-11
sympatric maize	r quintile	-0.110	0.011	-9.896	1.50E-21
sympatric maize	elevation*r quintile	0.074	0.005	14.399	8.27E-41
sympatric *mexicana*	intercept	0.270	0.035	7.710	3.35E-14
sympatric *mexicana*	elevation (km)	0.270	0.016	16.590	4.52E-54
sympatric *mexicana*	r quintile	0.119	0.014	8.309	3.57E-16
sympatric *mexicana*	elevation*r quintile	-0.045	0.007	-6.735	2.94E-11

Table 2.6. **Spearman's rank correlation between ancestry (f_4 ratio) and recombination rate (or coding bp per cM) quintiles in sympatric *mexicana***

group	feature	Spearman's ρ	2.5%	97.5%
sympatric *mexicana*	recombination rate (cM/Mb)	1.00	0.80	1.00
sympatric *mexicana*	coding bp per cM	-1.00	-1.00	-0.90

Table 2.7. **Spearman's rank correlation between mean *mexicana* local ancestry and recombination rate (or coding bp per cM) at 1 cM genomic window resolution.** Confidence intervals are constructed using the percentile method and 10,000 bootstrap replicates created by randomly re-sampling 1 cM windows within quintiles.

group	feature	Spearman's ρ	2.5%	97.5%
sympatric maize	recombination rate (cM/Mb)	0.011	-0.039	0.062
sympatric maize	coding bp per cM	0.014	-0.038	0.066
sympatric *mexicana*	recombination rate (cM/Mb)	-0.473	-0.512	-0.432
sympatric *mexicana*	coding bp per cM	0.339	0.292	0.384

Table 2.8. **Domestication genes and overlap with low introgression regions.**

gene	phenotype	references	maize v4 coordinates	minimum introgression in maize	minimum introgression in mexicana
zagl1	ear size	[**185**]	1:4959131-5014850	0.012*	0.264
gt1	prolificacy	[**186**]	1:23605801-23647370	0.004*	0.106
ZmSh1-1	seed shattering	[**190**]	1:228660490-228705551	0.356	0.038*
tb1	branching	[**179**, **180**, **181**]	1:270533676-270574776	0.001*	0.06*
zfl2	cob rank	[**192**, **193**, **194**]	2:12894091-12937068	0.113	0.271
pbf1	storage protein synthesis	[**196**]	2:158122366-158176919	0.092	0.084
ra2	inflorescence architecture	[**188**]	3:12138280-12179065	0.144	0.046*
ba1	plant architecture	[**195**]	3:185994629-186035264	0.33	0.154
su1	starch biosynthesis	[**184**]	4:43090569-43139167	0.445	0.017*
tga1	'naked' grains	[**182**, **183**]	4:46330597-46375118	0.036*	0.007*
bt2	starch biosynthesis	[**184**]	4:61295575-61341350	0.027*	0.016*
ZmSh1-5.1+ ZmSh1-5.2	seed shattering	[**190**]	5:16630307-16676707	0.105	0.175
sweet4c	sugar transport and seed size	[**187**]	5:130767030-130809864	0.014*	0.163
ae1	starch biosynthesis	[**184**]	5:172392995-172450415	0.436	0.059*
ra1	inflorescence architecture	[**188**, **189**]	7:113552410-113592937	0.006*	0.176

* lowest 5% introgression genomewide ('introgression desert')

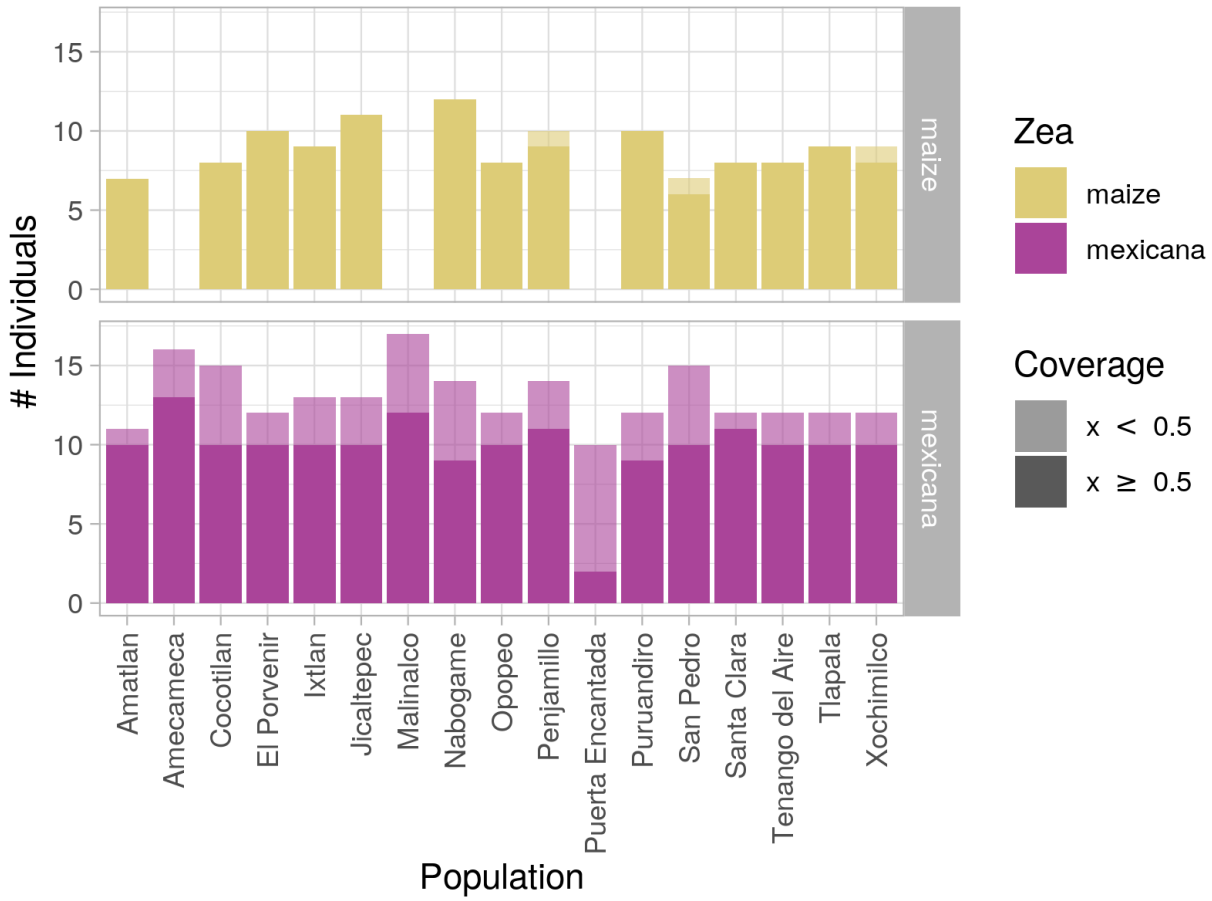


FIGURE 2.8. **Number of individuals sequenced per location**. Number of maize (top) and *mexicana* (bottom) individuals sequenced by this study with minimum 0.05x WGS coverage. Amecameca, Malinalco and Puerta Encantada have no paired maize samples and are used as a reference panel for *mexicana* ancestry. For sympatric maize and *mexicana*, only individuals meeting a more stringent 0.5x coverage threshold (shown in darker shading) are included in analyses based on local ancestry inference.

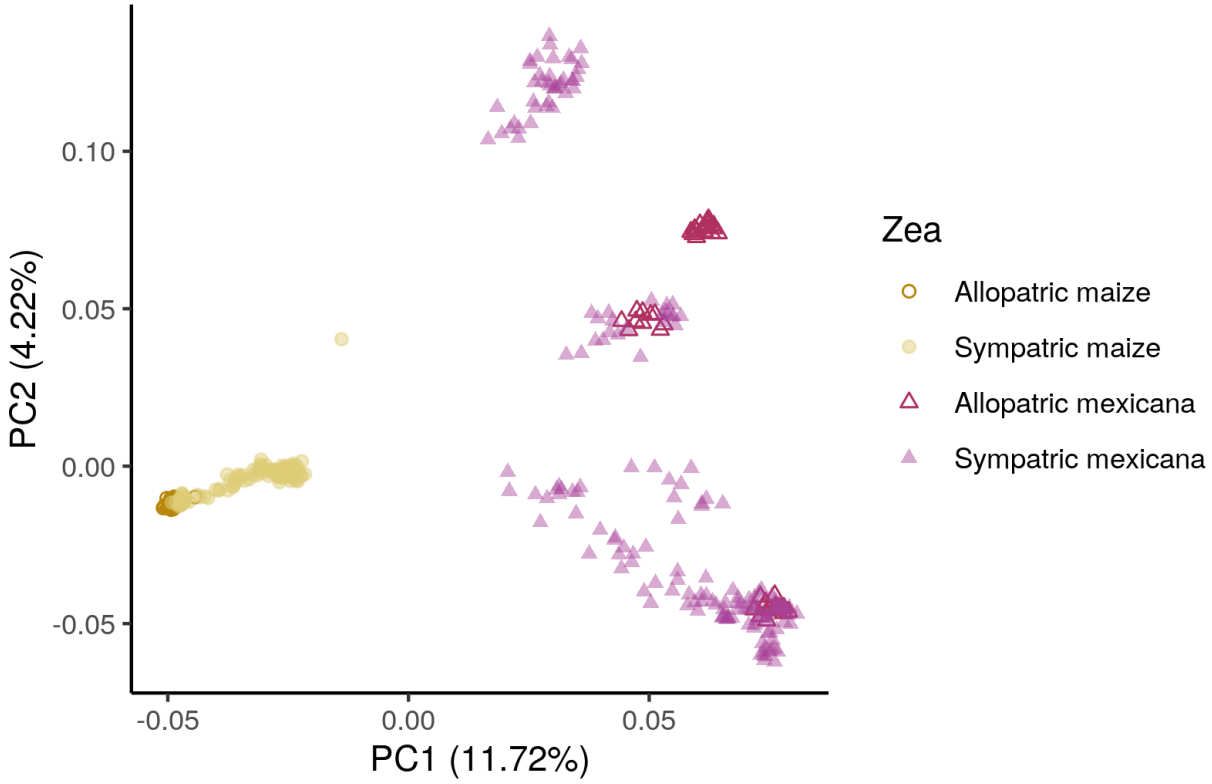


FIGURE 2.9. **PCA**. First and second principal components from the genomewide genetic covariance matrix relating sympatric and allopatric (reference) maize and *mexicana* individuals (PCAngsd). PC1 separates maize and *mexicana* subspecies while PC2 differentiates genetic clusters within *mexicana*.

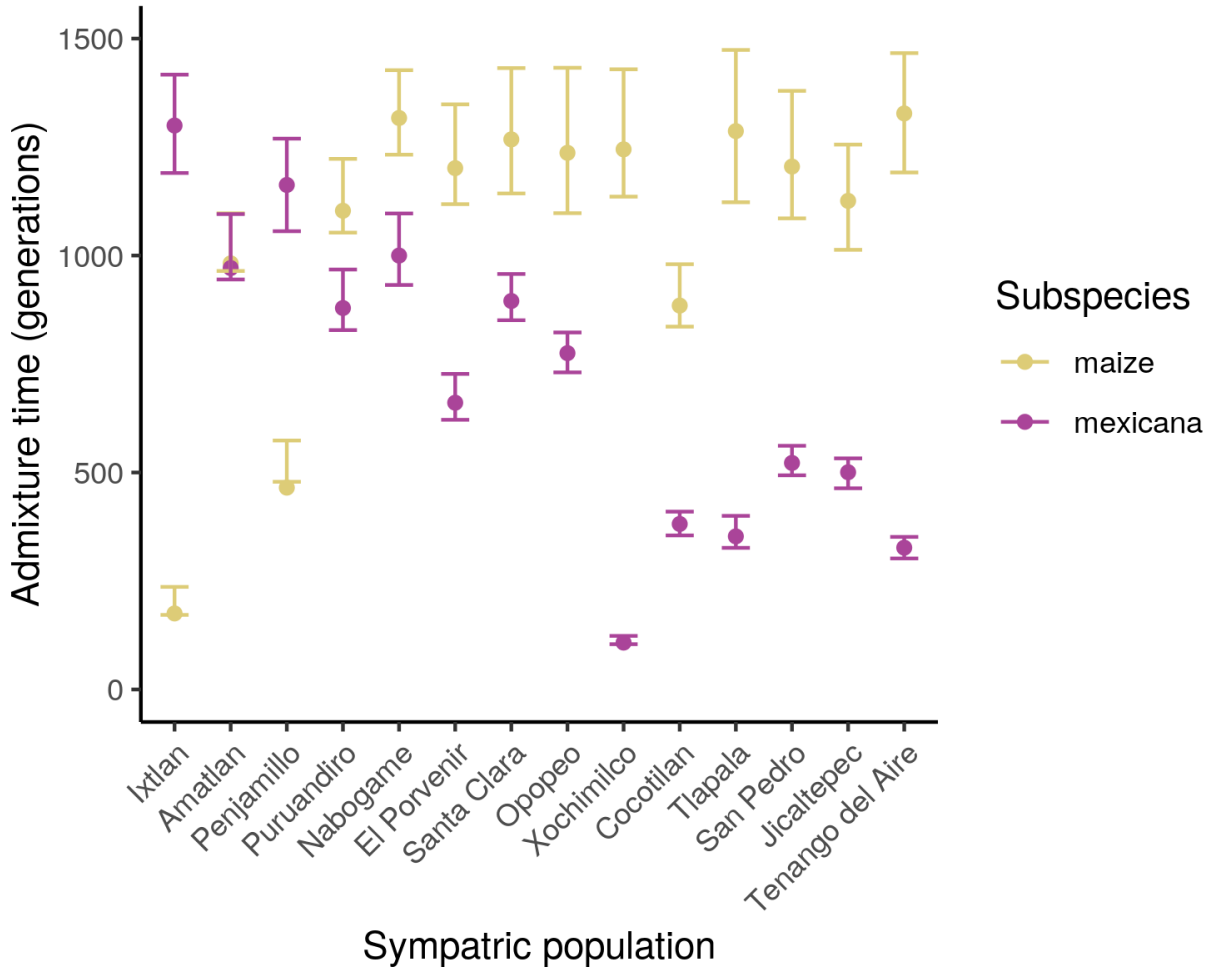


FIGURE 2.10. **Time since admixture**. Estimated generations since admixture under a single-pulse model for each sympatric maize and *mexicana* population, with 95% percentile confidence intervals based on 100 bootstrap samples of genomic blocks (1,000 SNPs per block). Estimates and bootstraps were produced during ancestry_hmm model fitting for local ancestry inference. Populations are ordered left to right by increasing elevation.

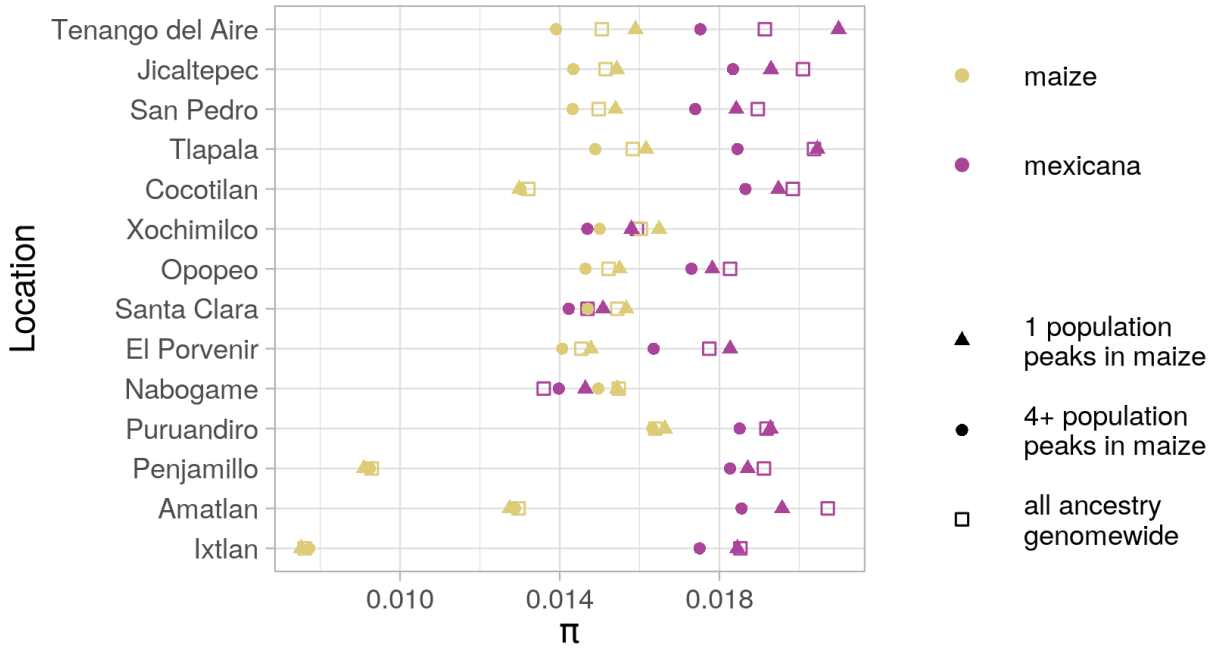


FIGURE 2.11. **Diversity (π) within *mexicana* ancestry** Each point summarises pairwise genetic diversity (π) for genomic regions with high-confidence homozygous *mexicana* ancestry, calculated separately for the maize and *mexicana* populations at each sampled location. Within-*mexicana* ancestry π is calculated and plotted separately for three subsets of the genome: introgression peaks ($>$ 2 s.d. above the mean) found in the focal maize population only, peaks shared between the focal maize and at least 3 other maize populations, and a genomewide estimate.

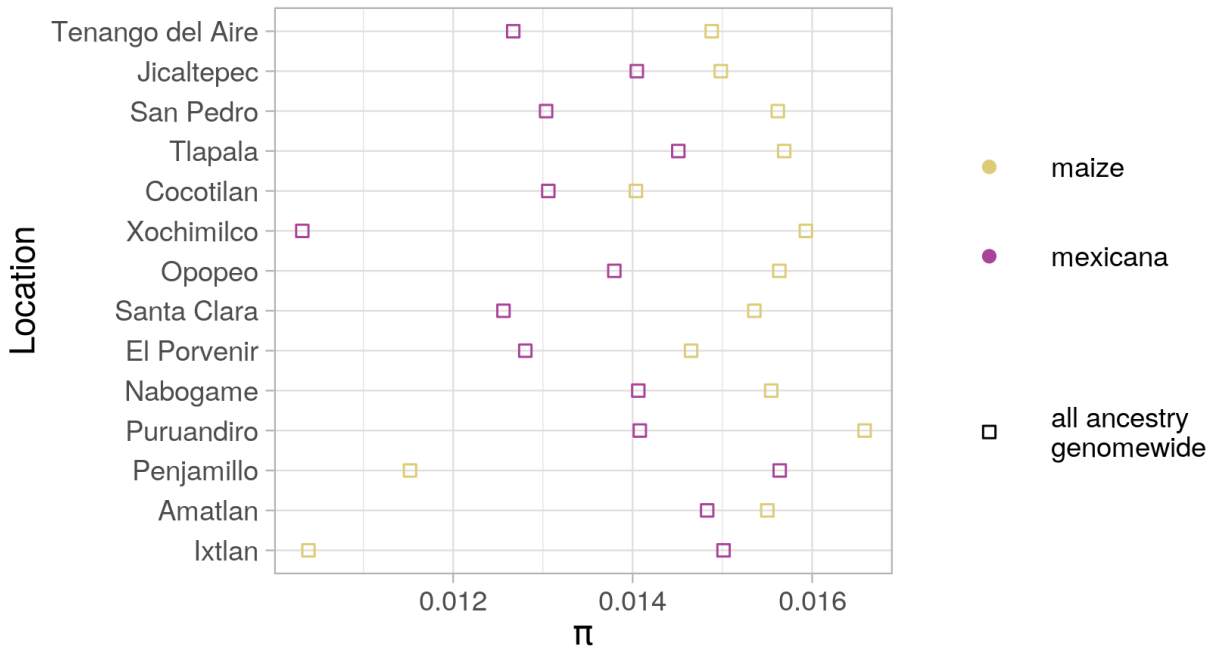


FIGURE 2.12. **Diversity (π) within maize ancestry** Each point summarises pairwise genetic diversity (π) for regions genomewide with high-confidence homozygous *mexicana* ancestry, calculated separately for the maize and *mexicana* populations at each sampled location.

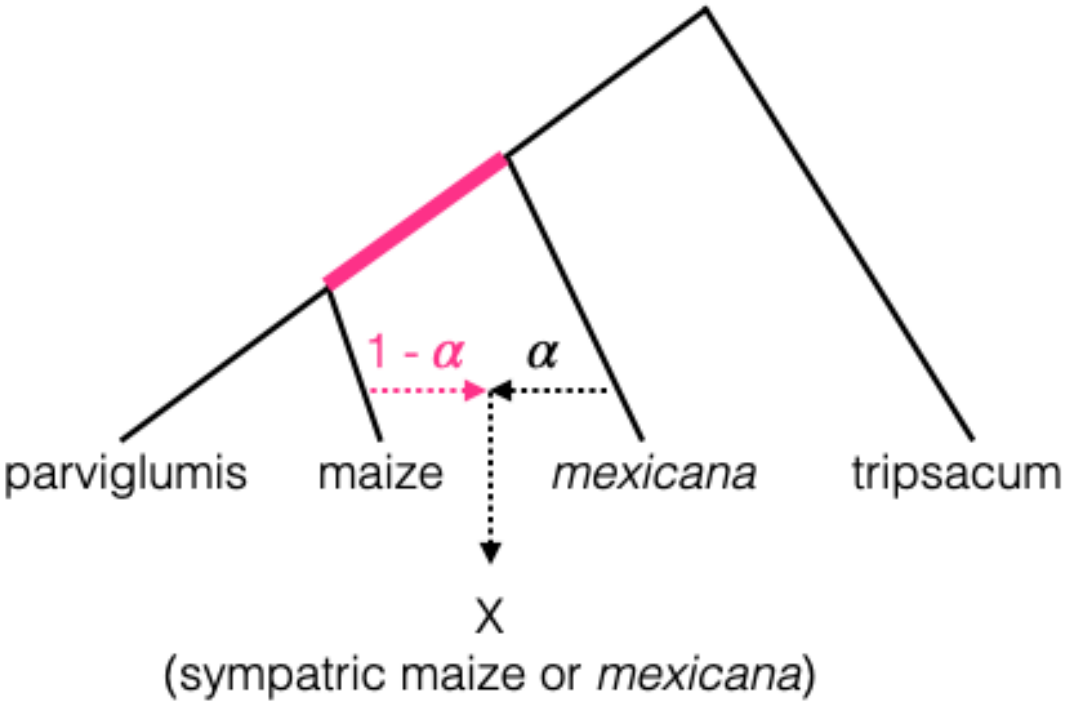


FIGURE 2.13. **Population tree** Phylogenetic tree assumed when estimating the ratio of f_4 statistics. The pink branch represents the shared drift between maize and parviglumis that is introduced to the focal sympatric population via admixture of proportion $1 - \alpha$. We used only plants from the Amecameca site in our *mexicana* reference group for this analysis because that site showed no evidence of previous admixture.

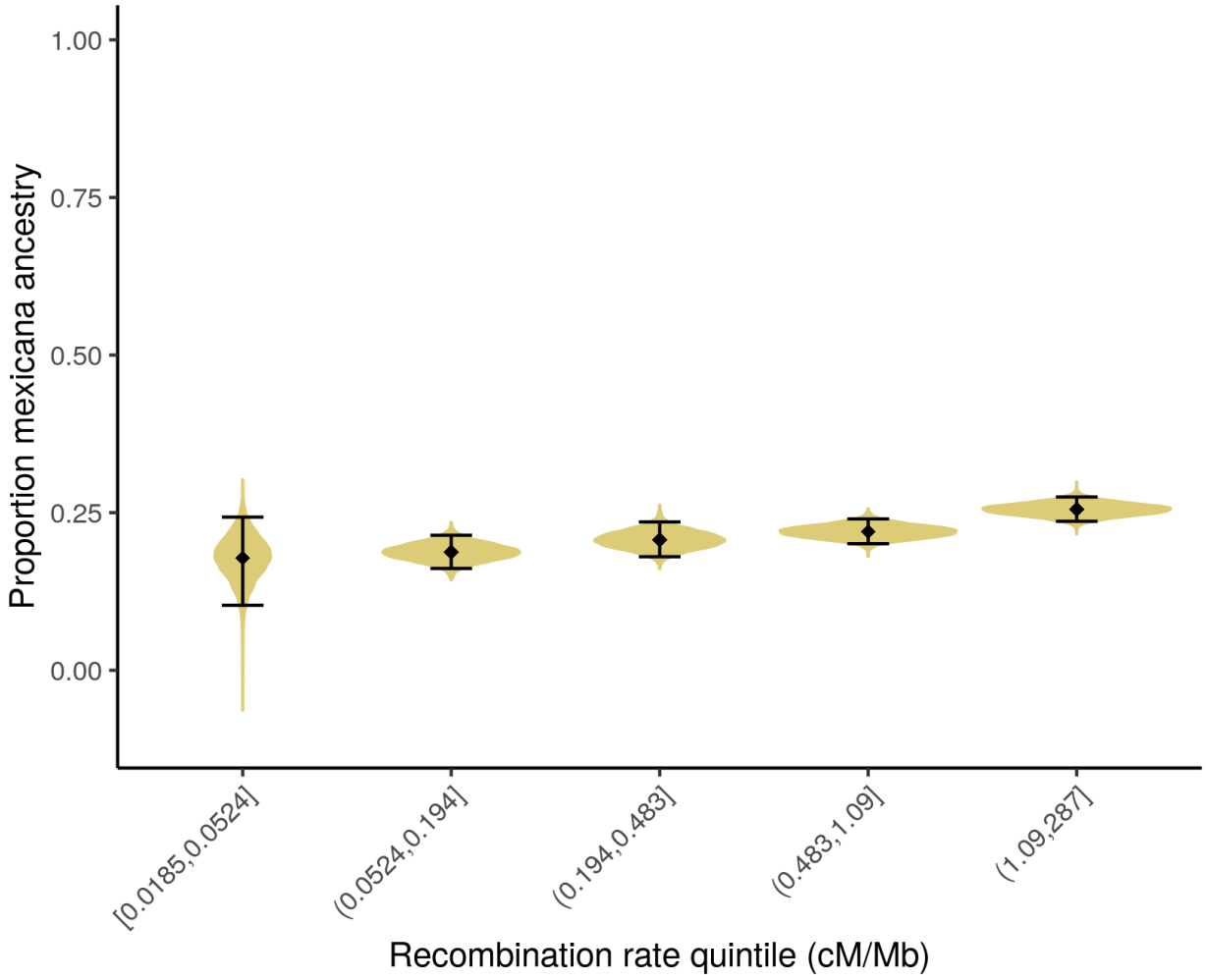


FIGURE 2.14. **f_4 ancestry in maize by recombination rate**. Estimated *mexicana* ancestry in sympatric maize landrace samples using f_4 ratio. Mean ancestry per recombination rate quintile and 95% percentile bootstrap confidence interval (n = 10,000) are depicted in black. Violin plots show the density of ancestry estimates for individual bootstraps re-sampled within quintiles.

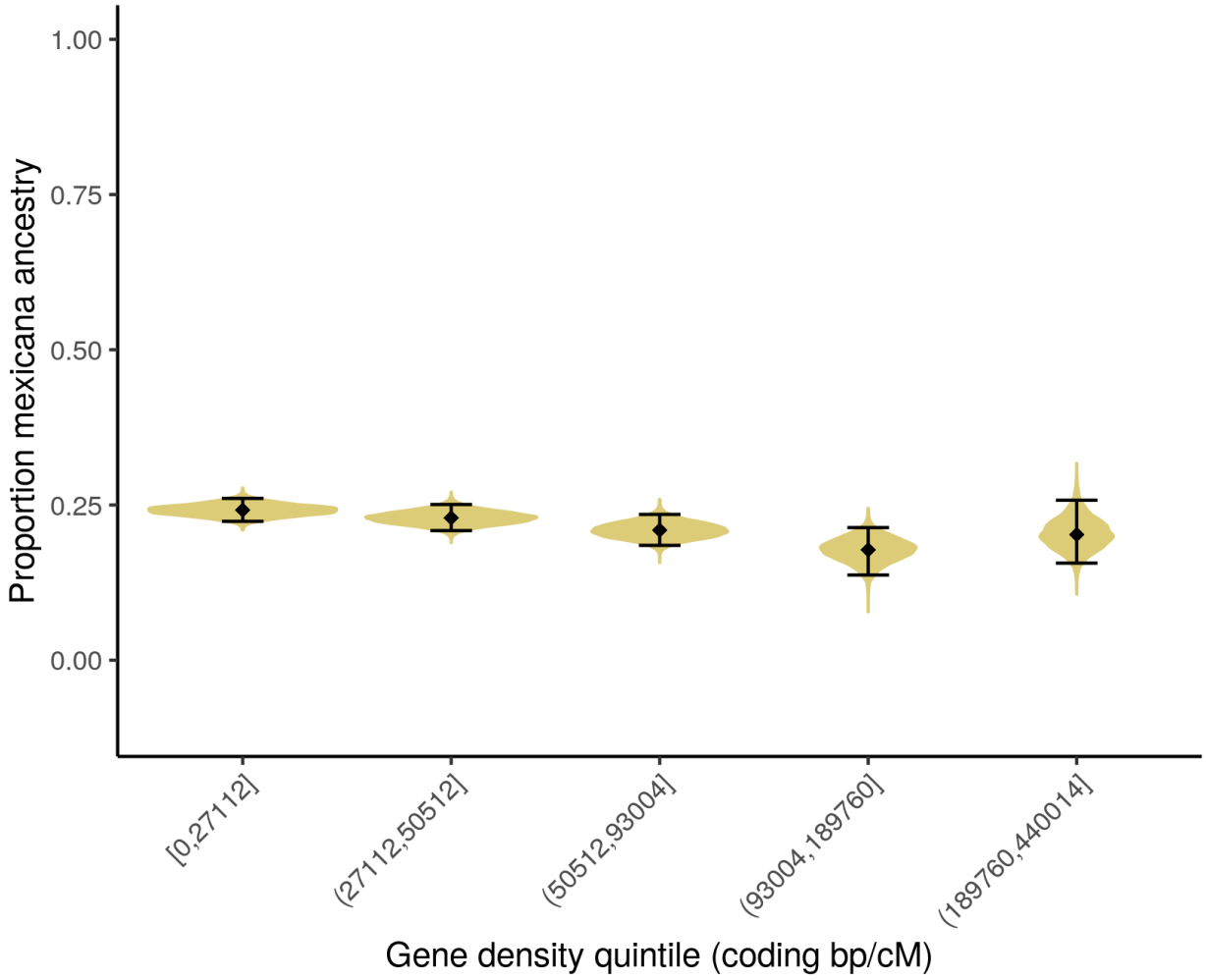


FIGURE 2.15. **f_4 ancestry in maize by coding bp per cM**. Estimated *mexicana* ancestry in sympatric maize landrace samples using f_4 ratio. Mean ancestry for each coding bp/cM quintile and 95% percentile bootstrap confidence interval (n = 10,000) are depicted in black. Violin plots show the density of ancestry estimates for individual bootstraps re-sampled within quintiles.

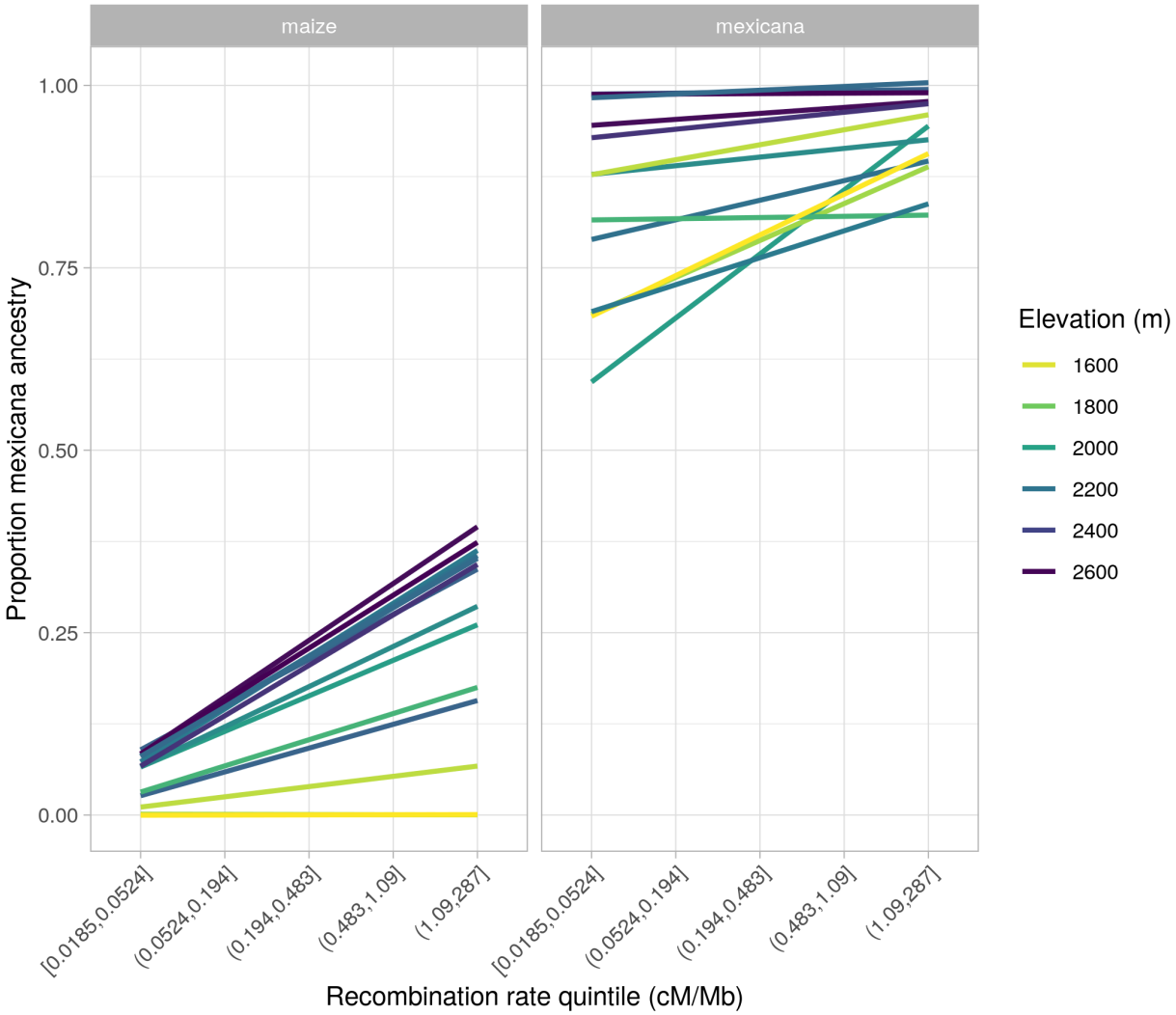


FIGURE 2.16. ***Mexicana* ancestry across recombination quintiles by elevation**. Estimated linear relationship between proportion *mexicana* ancestry (NGSAdmix) and recombination rate quintile for each sympatric population, colored by elevation. Linear models were fit using *lm()* in R on individuals' ancestry estimates per quintile.

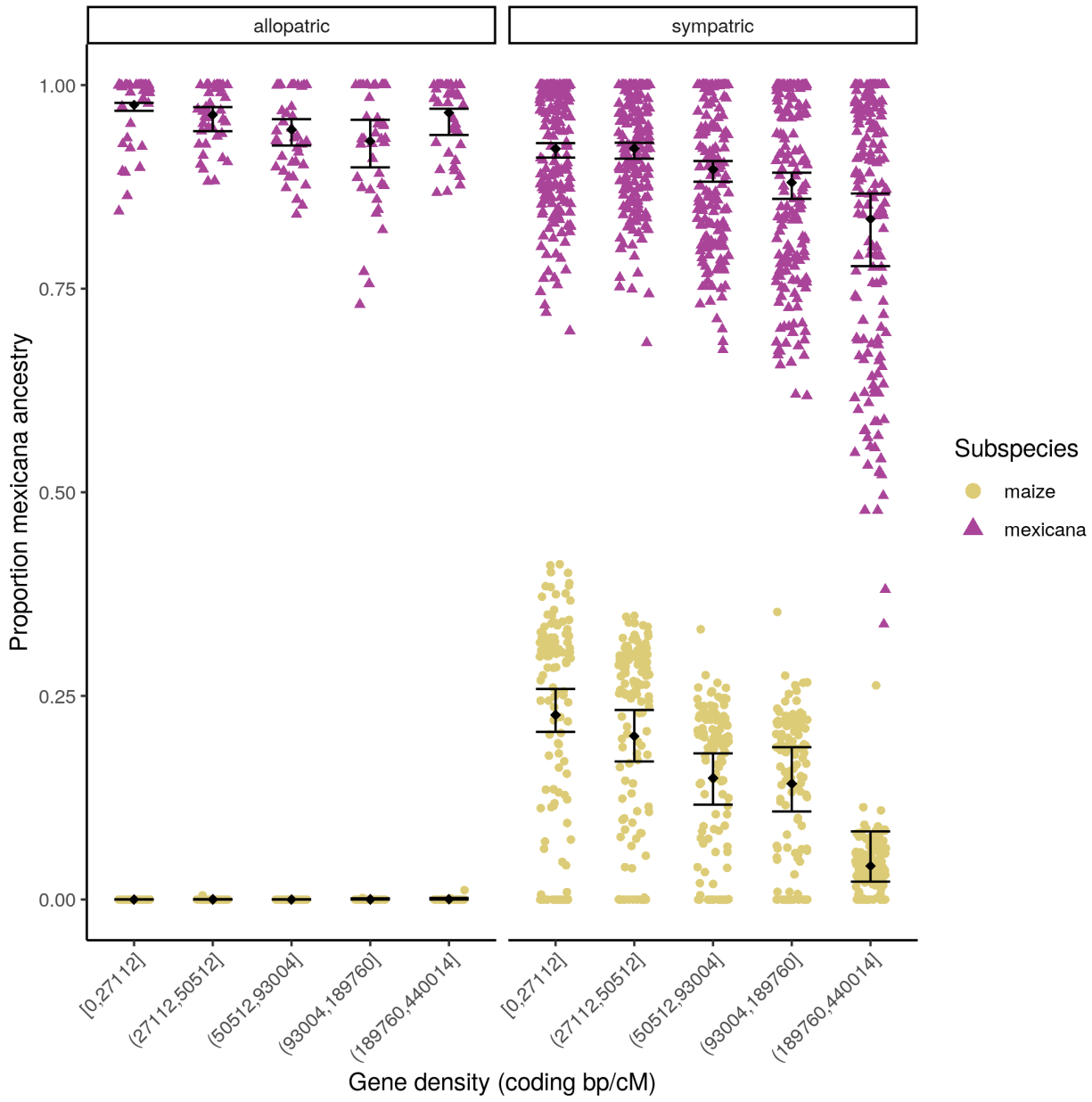


FIGURE 2.17. ***Mexicana* ancestry by coding bp per cM**. Inferred *mexicana* ancestry in allopatric reference populations (left) and sympatric maize and *mexicana* populations (right) using NGSAdmix (K=2) by coding density quintiles. Group mean and 95% percentile bootstrap confidence interval (n = 100) are depicted in black. Ancestry estimates for each individual are shown as points, colored by *Zea* subspecies, and points are jittered for better visualization.

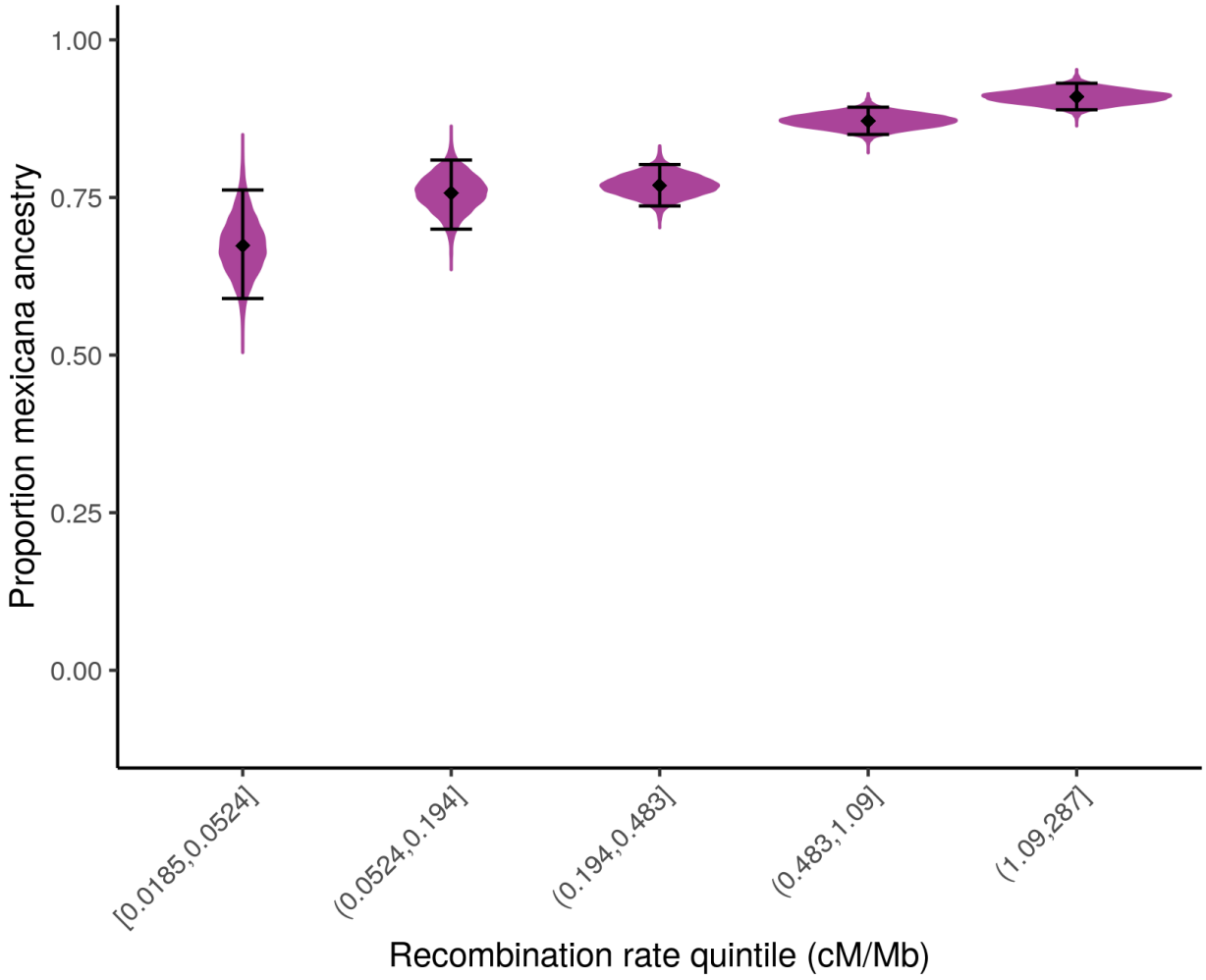


FIGURE 2.18. f_4 **ancestry in maize by recombination rate**. Estimated *mexicana* ancestry in sympatric mexicana samples using f_4 ratio. Mean ancestry per recombination rate quintile and 95% percentile bootstrap confidence interval (n = 10,000) are depicted in black. Violin plots show the density of ancestry estimates for individual bootstraps re-sampled within quintiles.

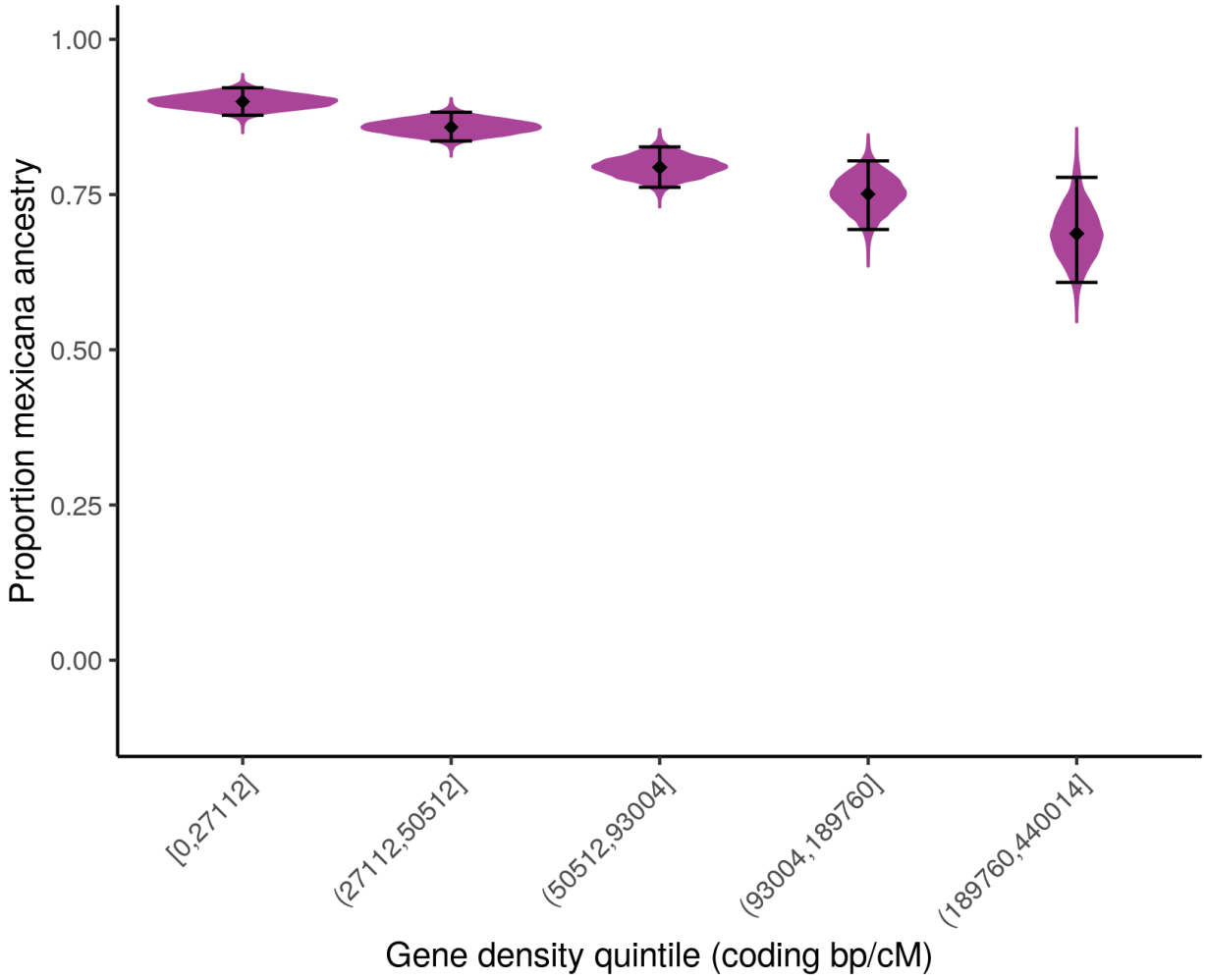


FIGURE 2.19. f_4 **ancestry in mexicana by coding bp per cM**. Estimated *mexicana* ancestry in sympatric mexicana samples using f_4 ratio. Mean ancestry for each coding bp/cM quintile and 95% percentile bootstrap confidence interval (n = 10,000) are depicted in black. Violin plots show the density of ancestry estimates for individual bootstraps re-sampled within quintiles.

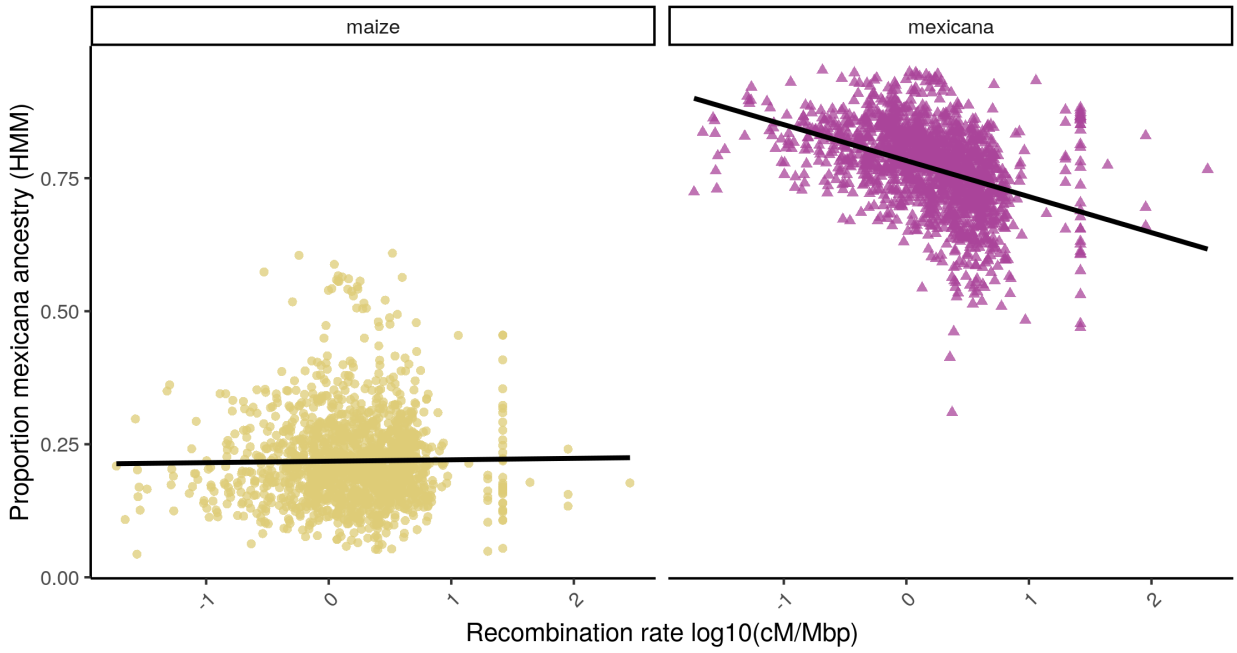


FIGURE 2.20. **Local *mexicana* ancestry in 1 cM windows by recombination rate**. Estimated *mexicana* ancestry in sympatric maize and *mexicana* samples using ancestry_hmm. Each point is a 1 cM genomic window and the line shows the best linear model fit for mean *mexicana* ancestry by recombination rate on a log scale.

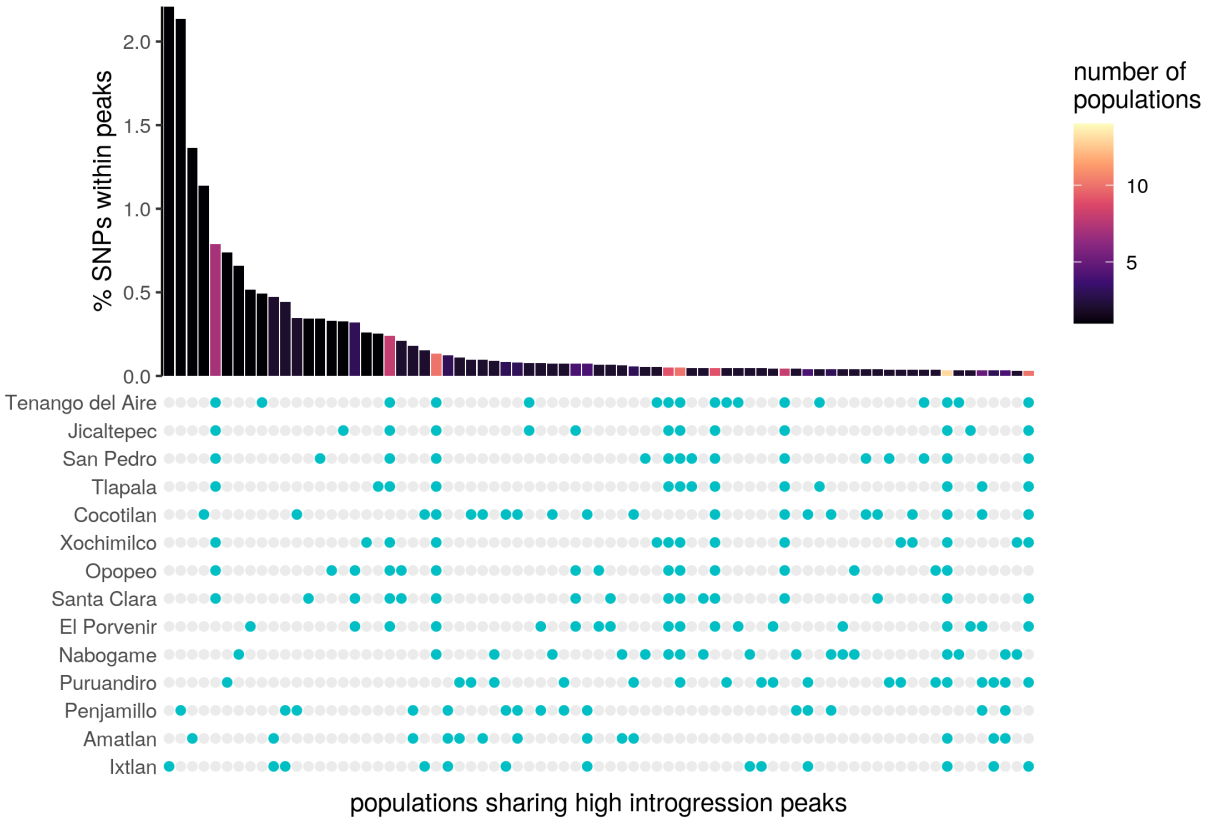


FIGURE 2.21. **High introgression peaks shared across sympatric maize populations** Here we show the 75 most common combinations of populations that share ancestry peaks (introgressed ancestry $>$ 2 s.d. above each population's mean ancestry). Bar height represents the percent of SNPs genomewide within peaks shared by the populations highlighted in blue below. Populations are ordered from high (top) to low elevation. See 2.22 for sympatric *mexicana* equivalent visualization.

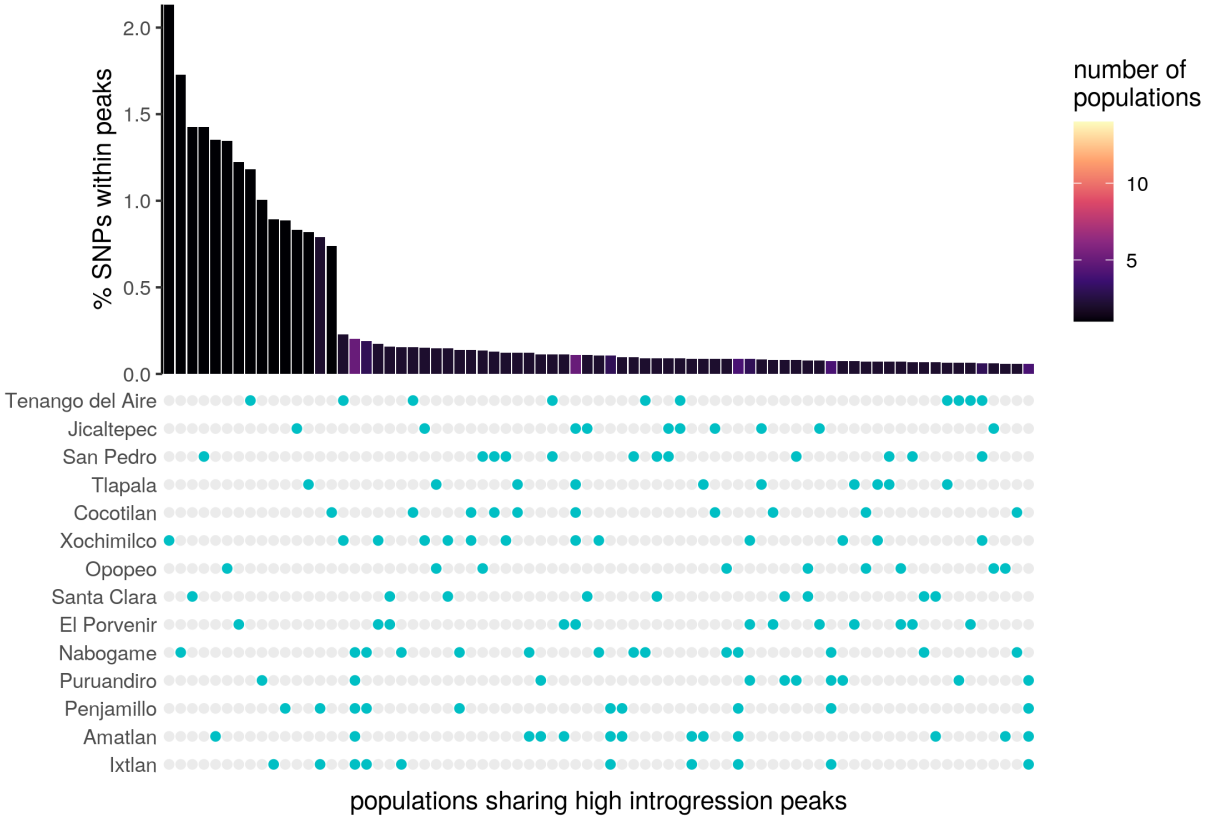


FIGURE 2.22. **High introgression peaks shared across sympatric *mexicana* populations** Here we show the 75 most common combinations of populations that share ancestry peaks (introgressed ancestry $>$ 2 s.d. above each population's mean ancestry). Bar height represents the percent of SNPs genomewide within peaks shared by the populations highlighted in blue below. Populations are ordered from high (top) to low elevation.

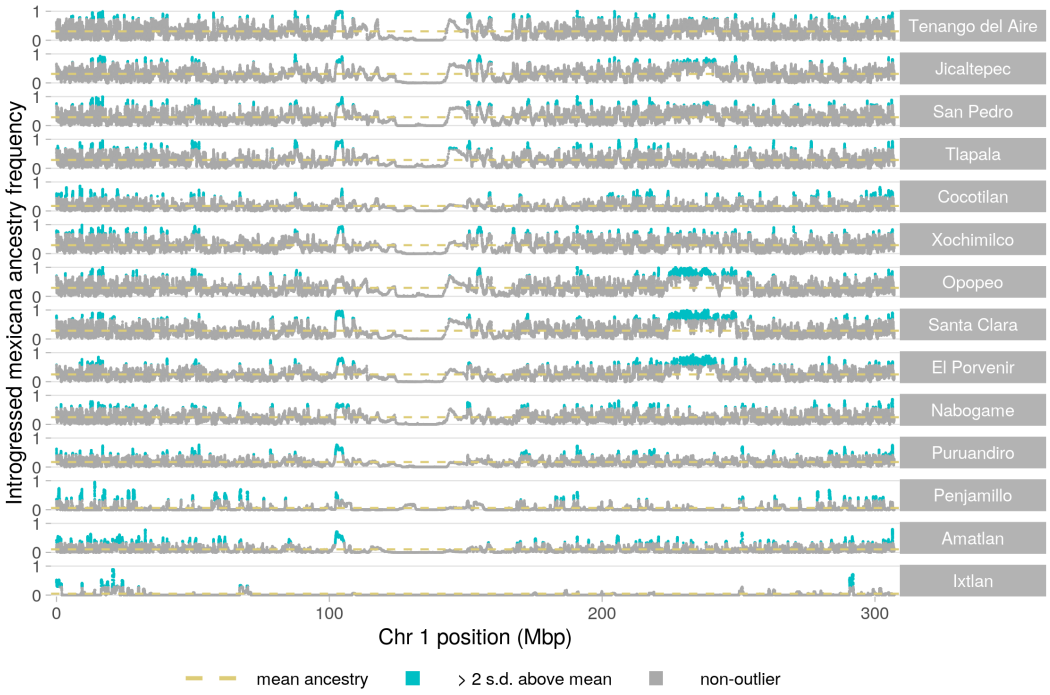


FIGURE 2.23. **Introgression in maize landrace populations across chromosome 1**

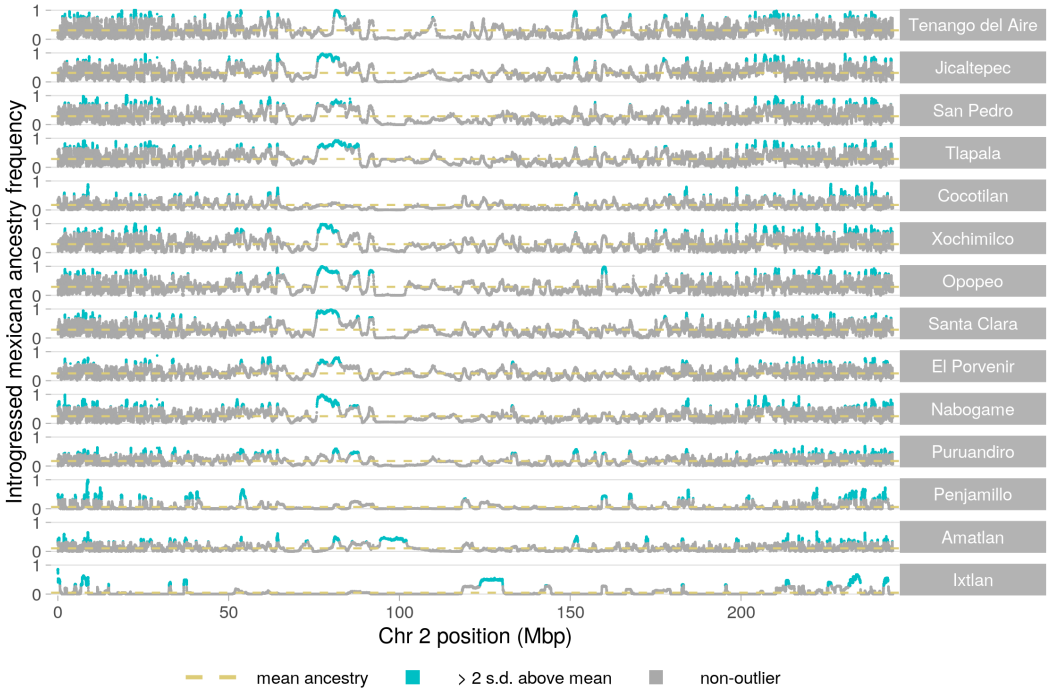


FIGURE 2.24. **Introgression in maize landrace populations across chromosome 2**

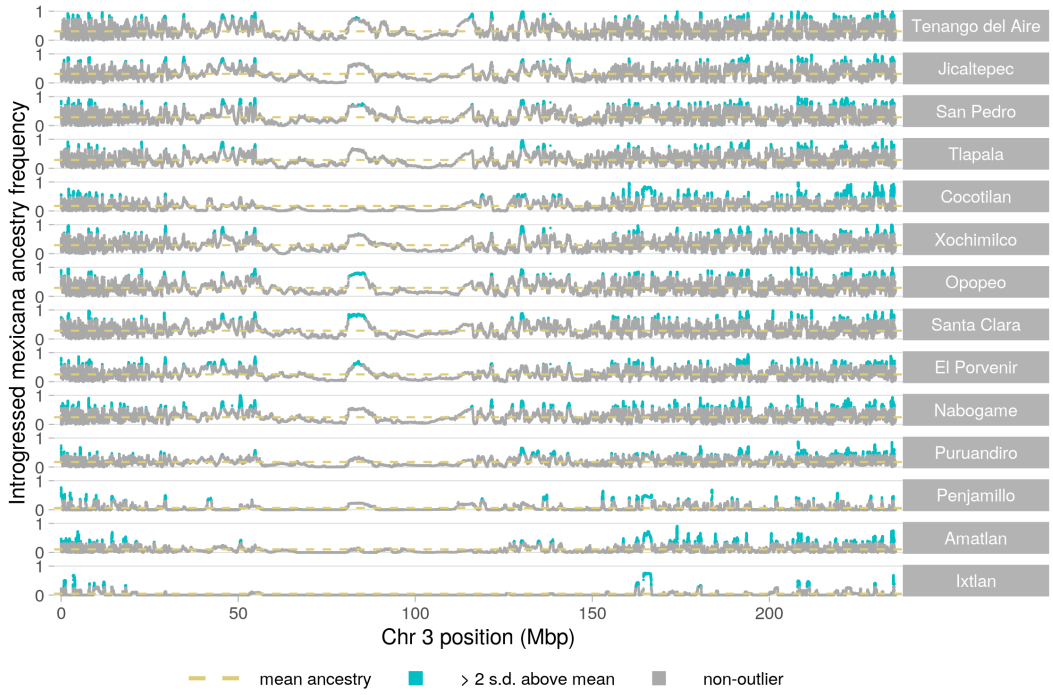


FIGURE 2.25. **Introgression in maize landrace populations across chromosome 3**

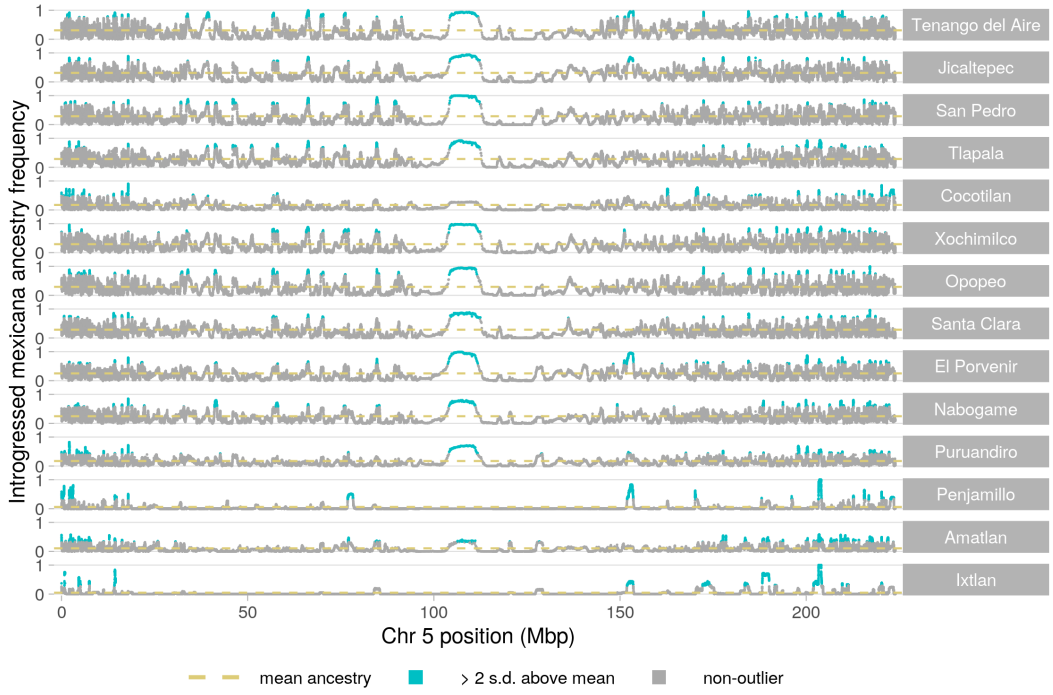


FIGURE 2.26. **Introgression in maize landrace populations across chromosome 5**

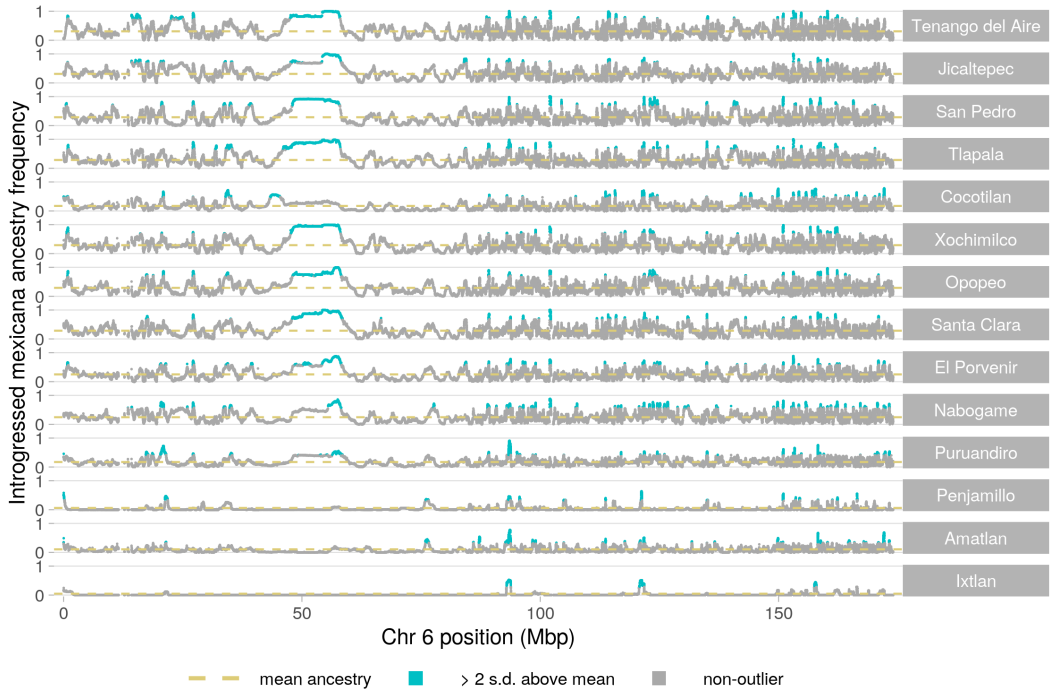


FIGURE 2.27. **Introgression in maize landrace populations across chromosome 6**

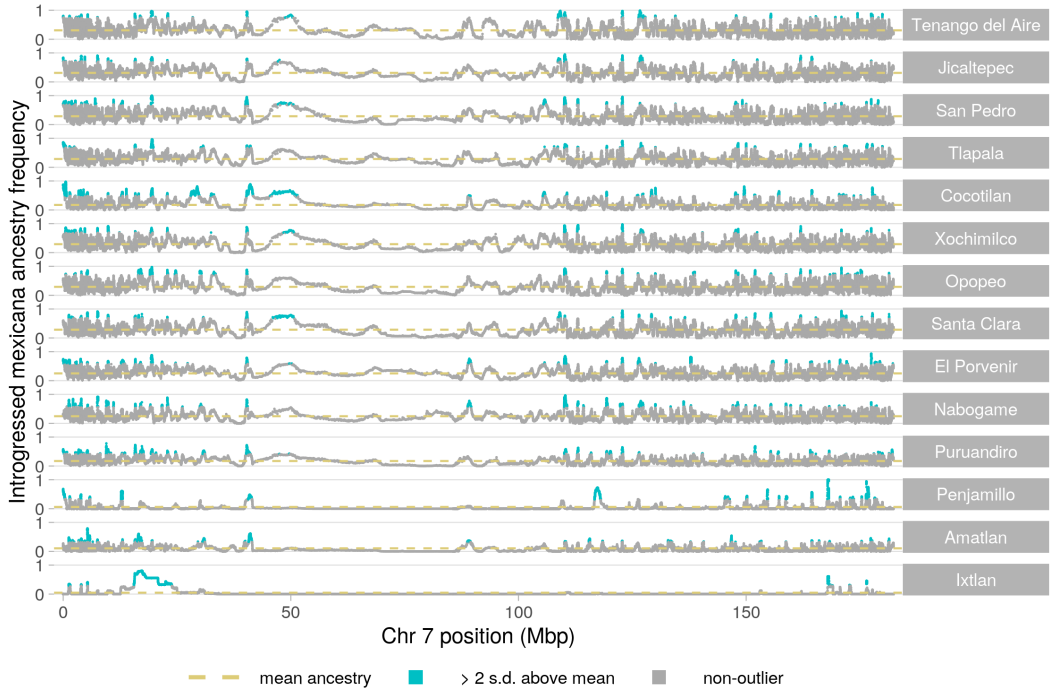


FIGURE 2.28. **Introgression in maize landrace populations across chromosome 7**

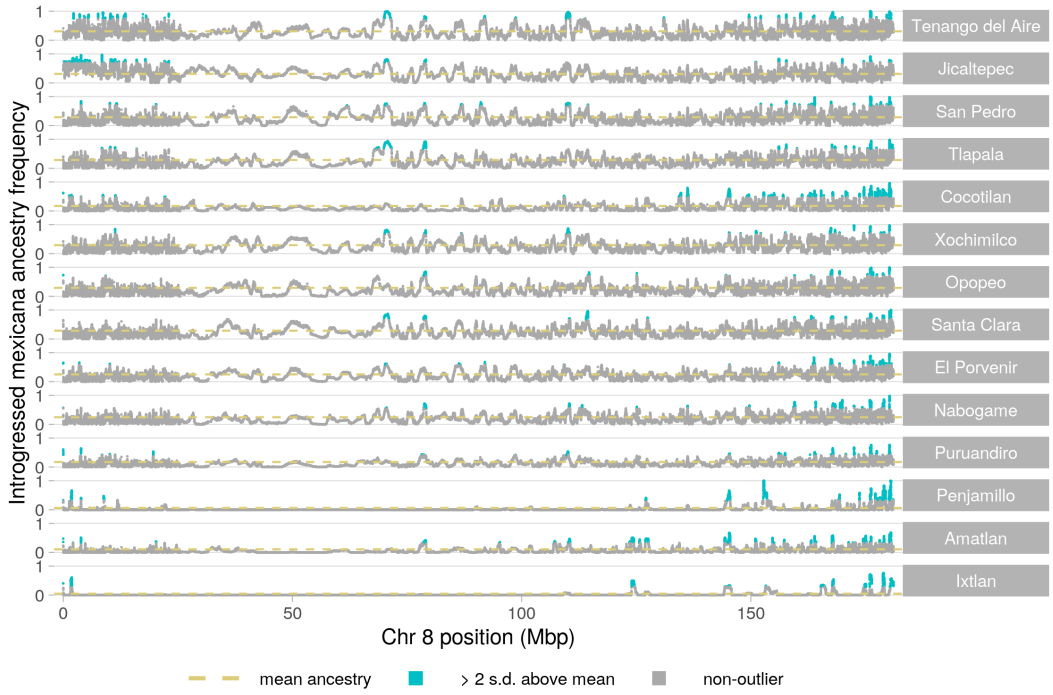


FIGURE 2.29. **Introgression in maize landrace populations across chromosome 8**

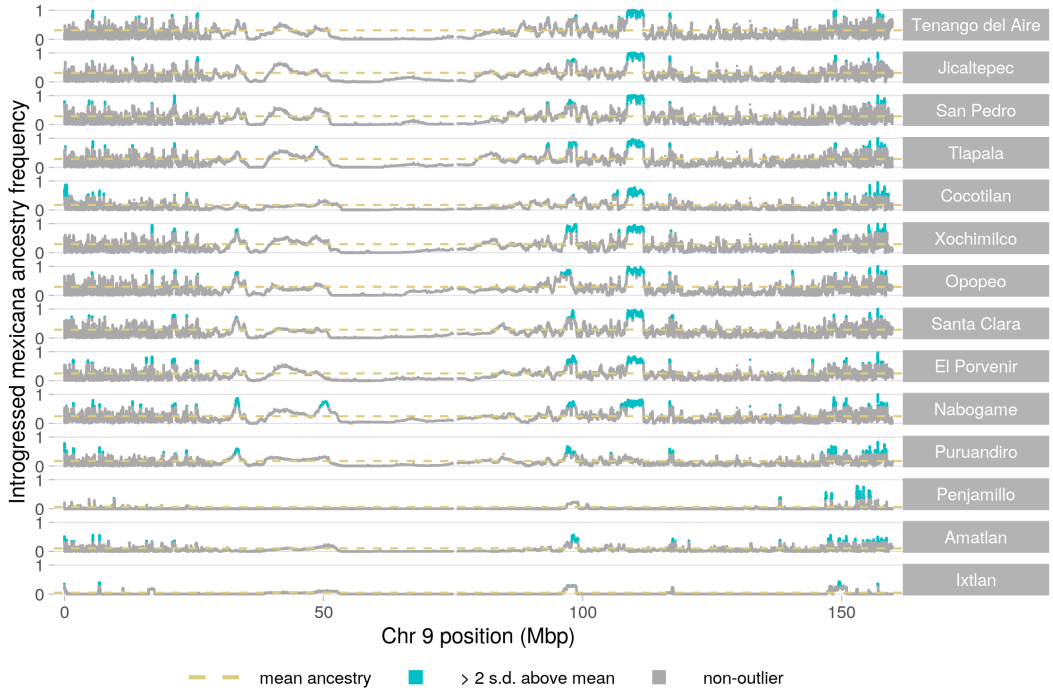


FIGURE 2.30. **Introgression in maize landrace populations across chromosome 9**

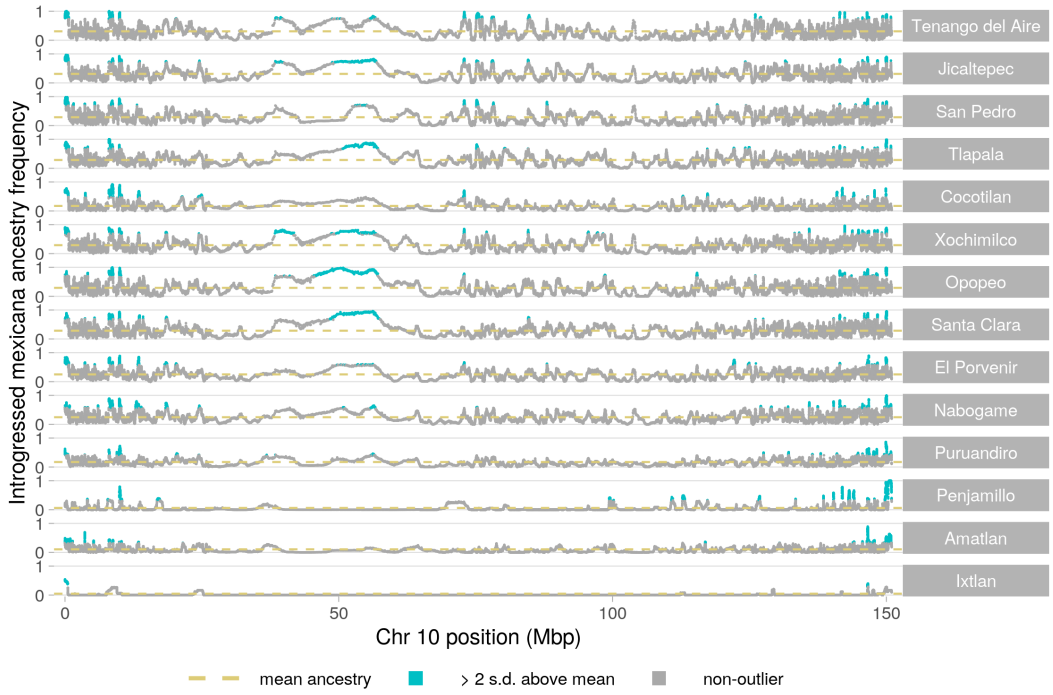


FIGURE 2.31. **Introgression in maize landrace populations across chromosome 10**

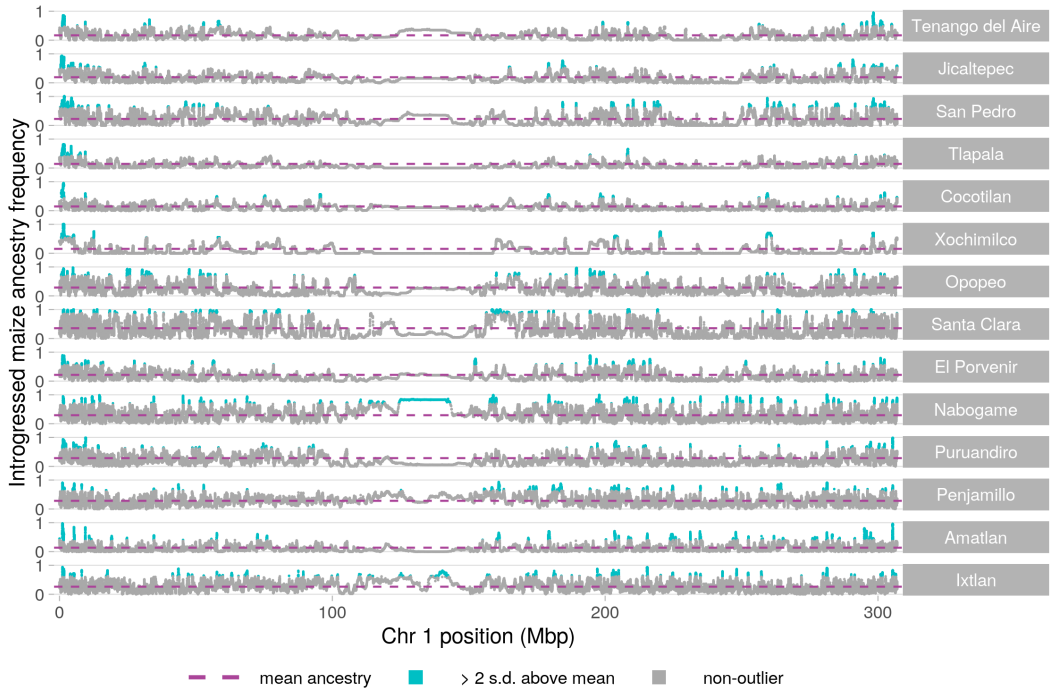


FIGURE 2.32. **Introgression in *mexicana* populations across chromosome 1**

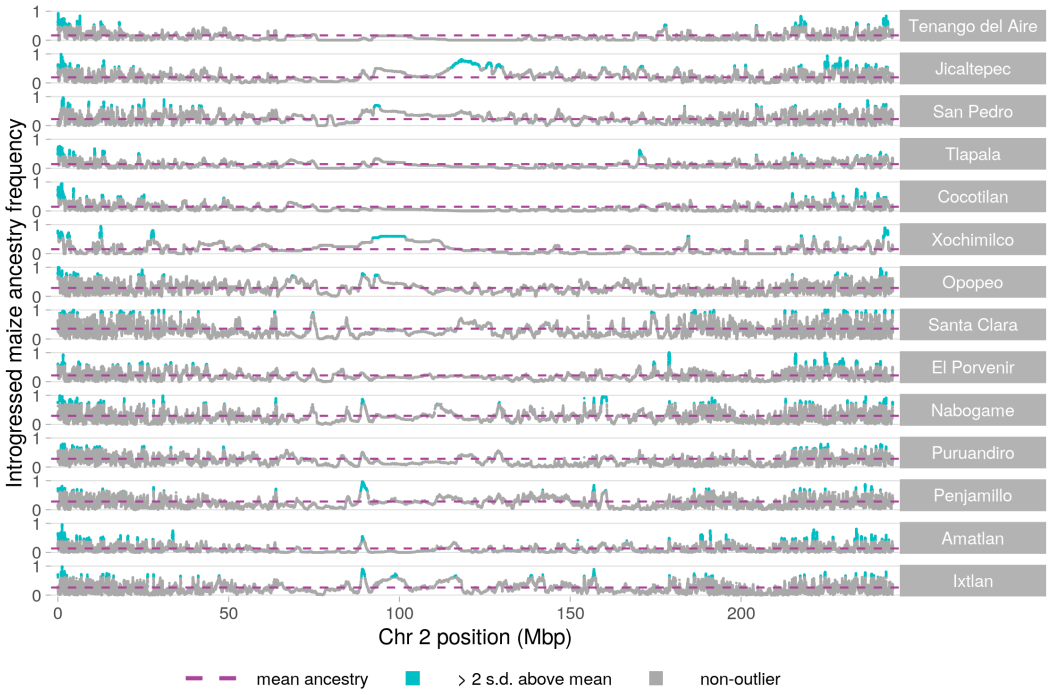


FIGURE 2.33. **Introgression in *mexicana* populations across chromosome 2**

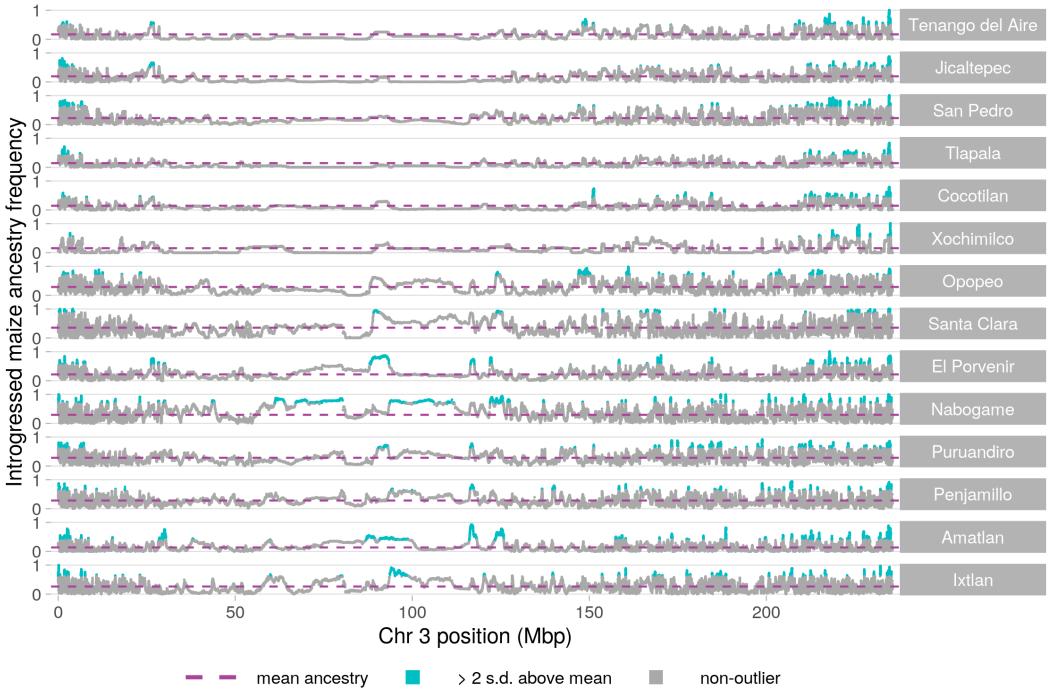


FIGURE 2.34. **Introgression in *mexicana* populations across chromosome 3**

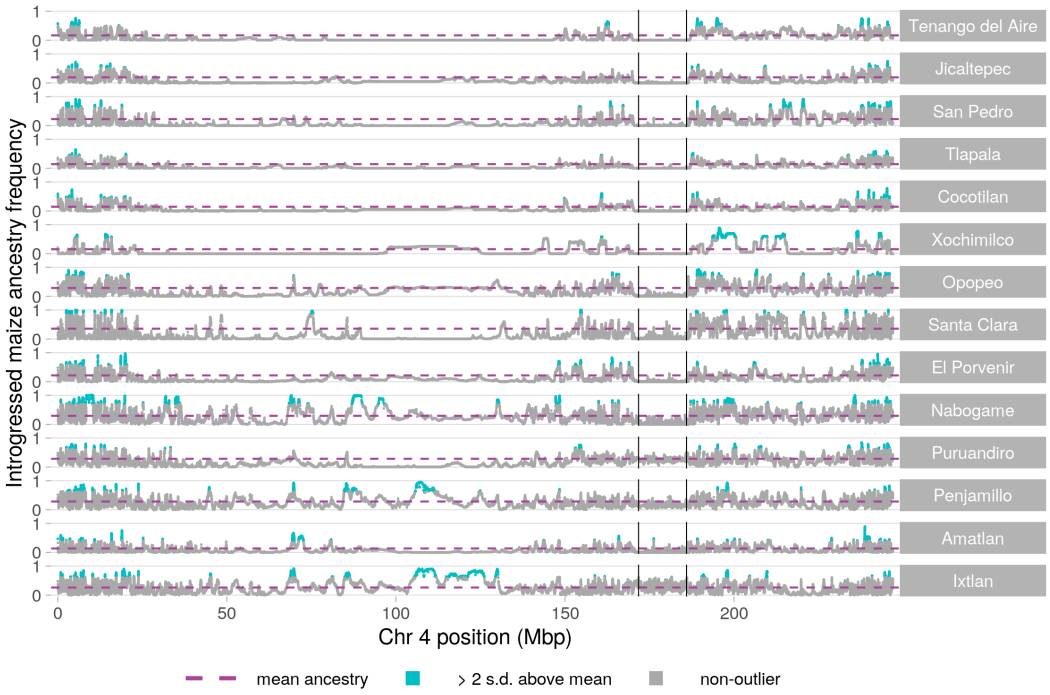


FIGURE 2.35. **Introgression in *mexicana* populations across chromosome 4** Vertical lines indicate the coordinates for *Inv4m*.

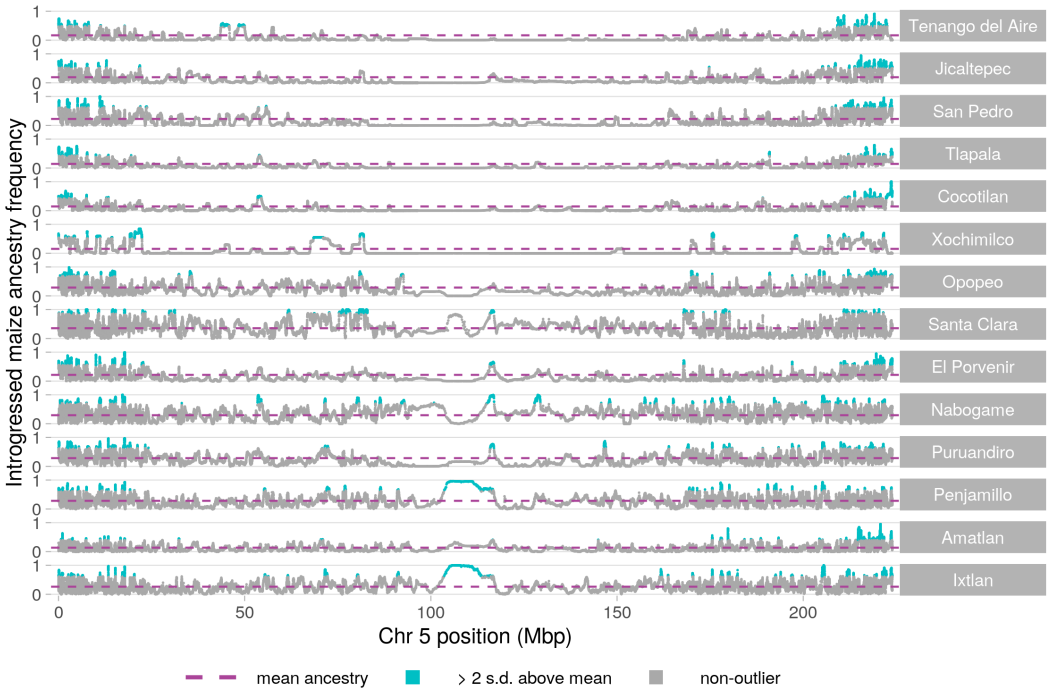


FIGURE 2.36. **Introgression in *mexicana* populations across chromosome 5**

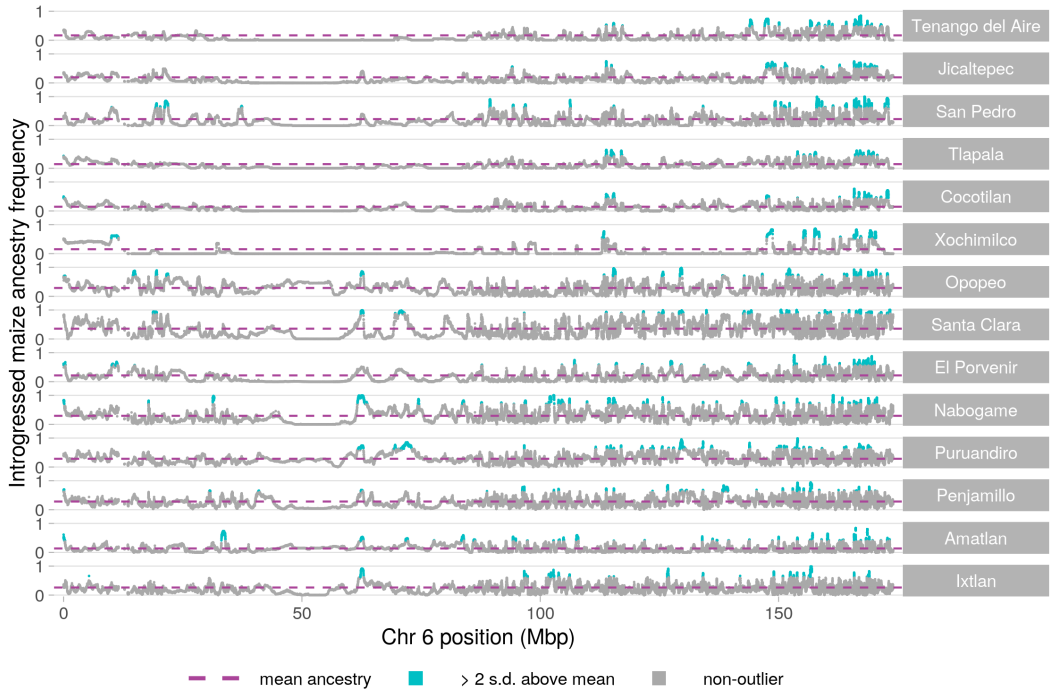


FIGURE 2.37. **Introgression in *mexicana* populations across chromosome 6**

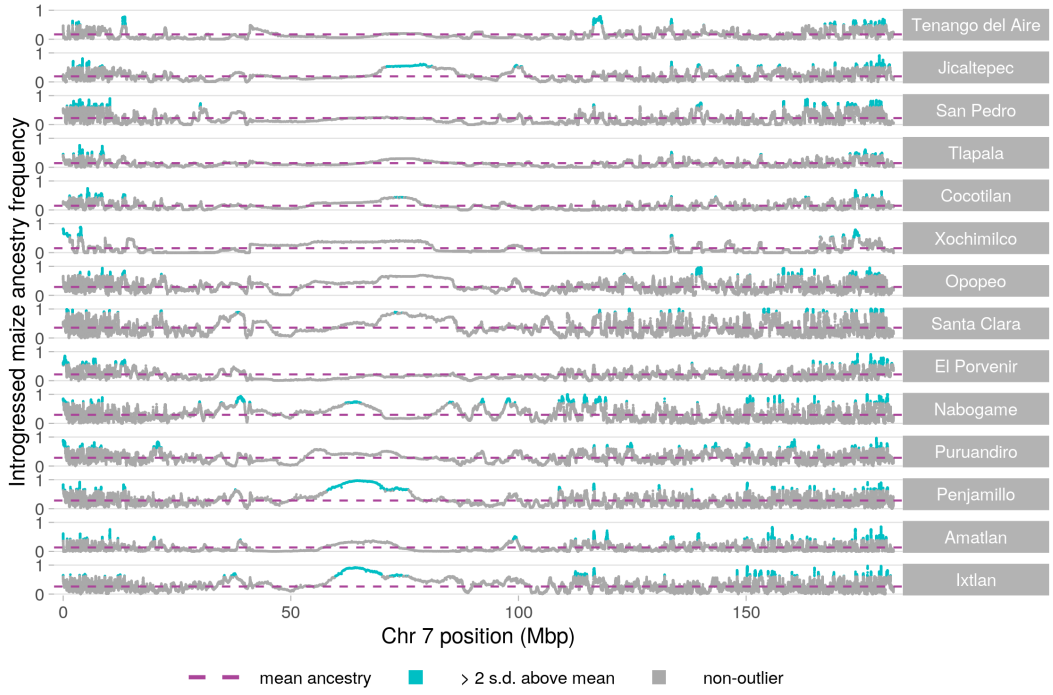


FIGURE 2.38. **Introgression in *mexicana* populations across chromosome 7**

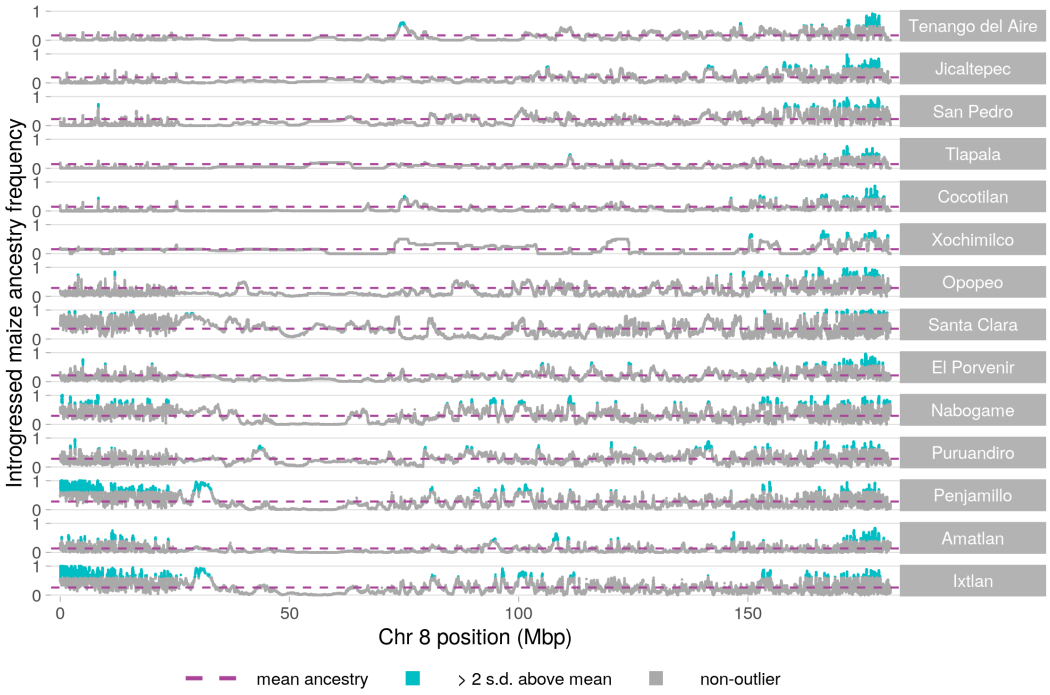


FIGURE 2.39. **Introgression in *mexicana* populations across chromosome 8**

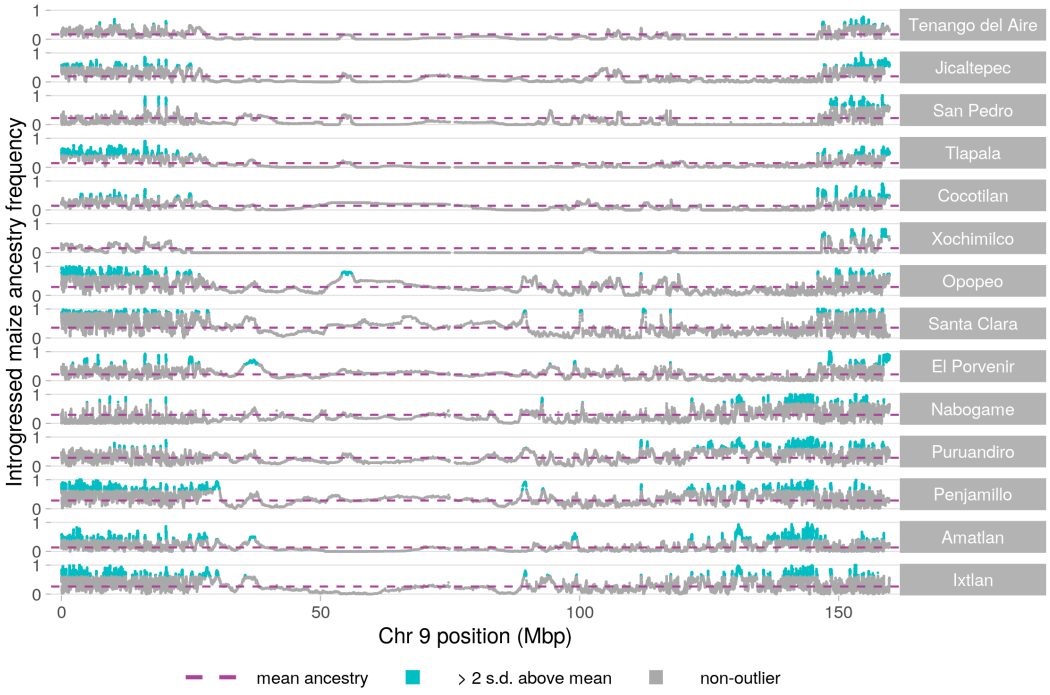


FIGURE 2.40. **Introgression in *mexicana* populations across chromosome 9**

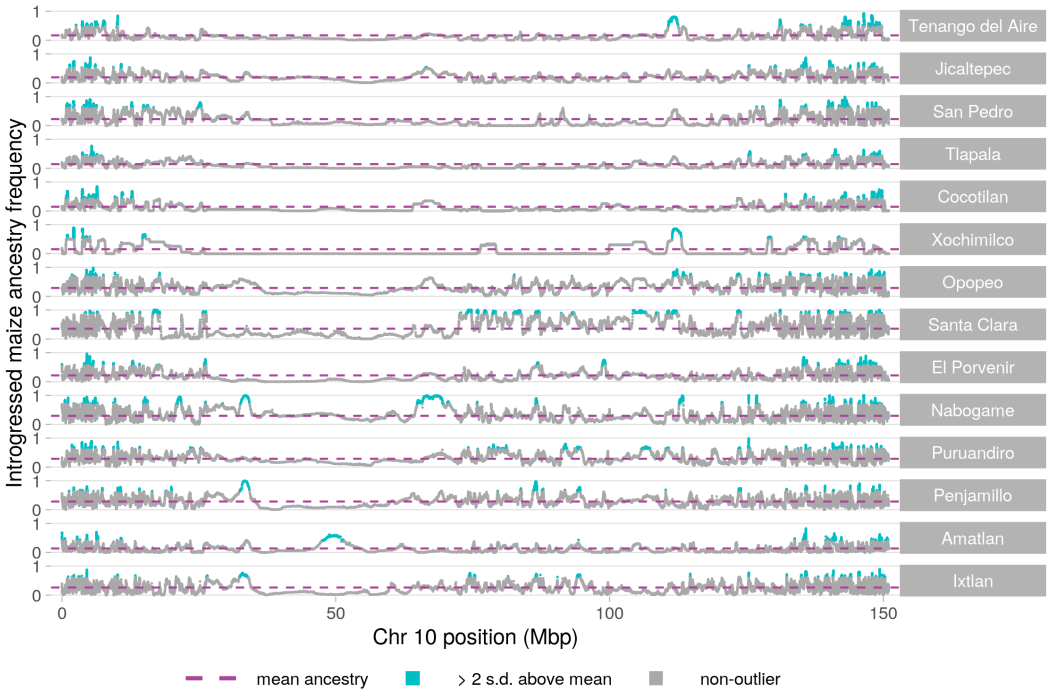


FIGURE 2.41. **Introgression in *mexicana* populations across chromosome 10**

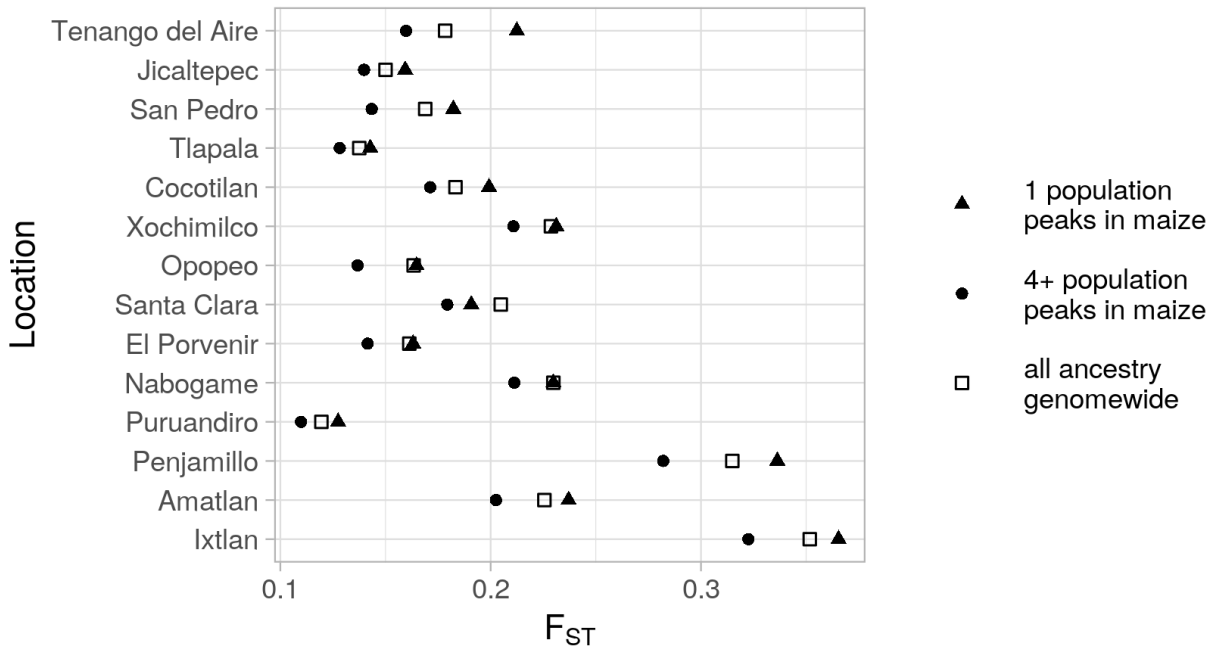


FIGURE 2.42. **Differentiation (F_{ST}) between introgressed ancestry tracts and local *mexicana*** Each point summarises F_{ST} between *mexicana* ancestry tracts within a focal maize population and *mexicana* ancestry tracts within the local *mexicana* population sampled at the same site. Within-*mexicana* ancestry F_{ST} is presented separately for three subsets of the genome: introgression peaks found in the focal maize population only, peaks shared between the focal maize and at least 3 other maize populations, and a genomewide estimate.

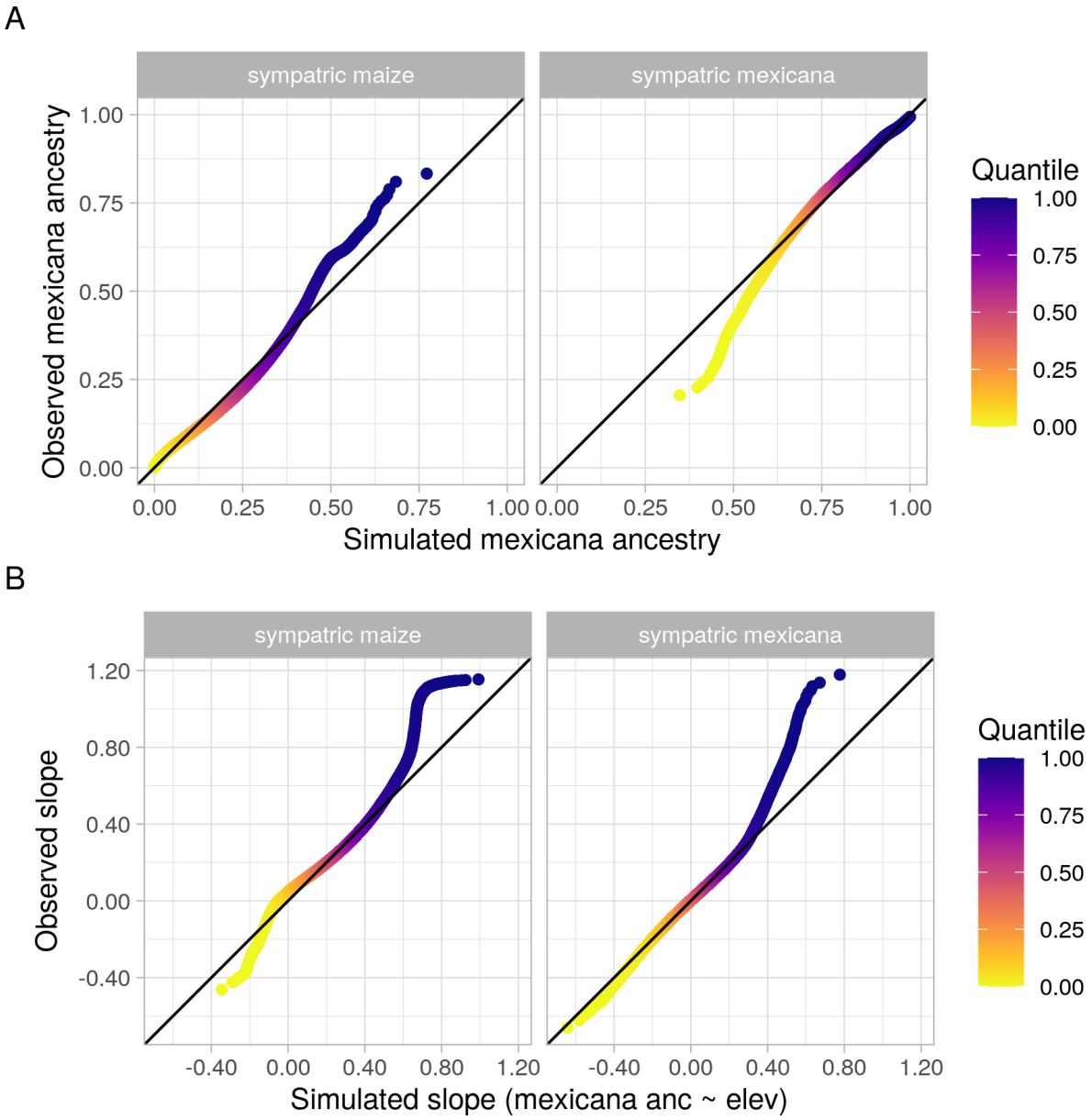


FIGURE 2.43. **Quantile comparison of observed data vs. MVN normal null model** (A) QQ-plot of simulated vs. observed mean ancestry at individual loci across all sympatric populations. (B) QQ-plot of simulated vs. observed slopes from the linear model *mexicana* ancestry ~ elevation at individual loci.

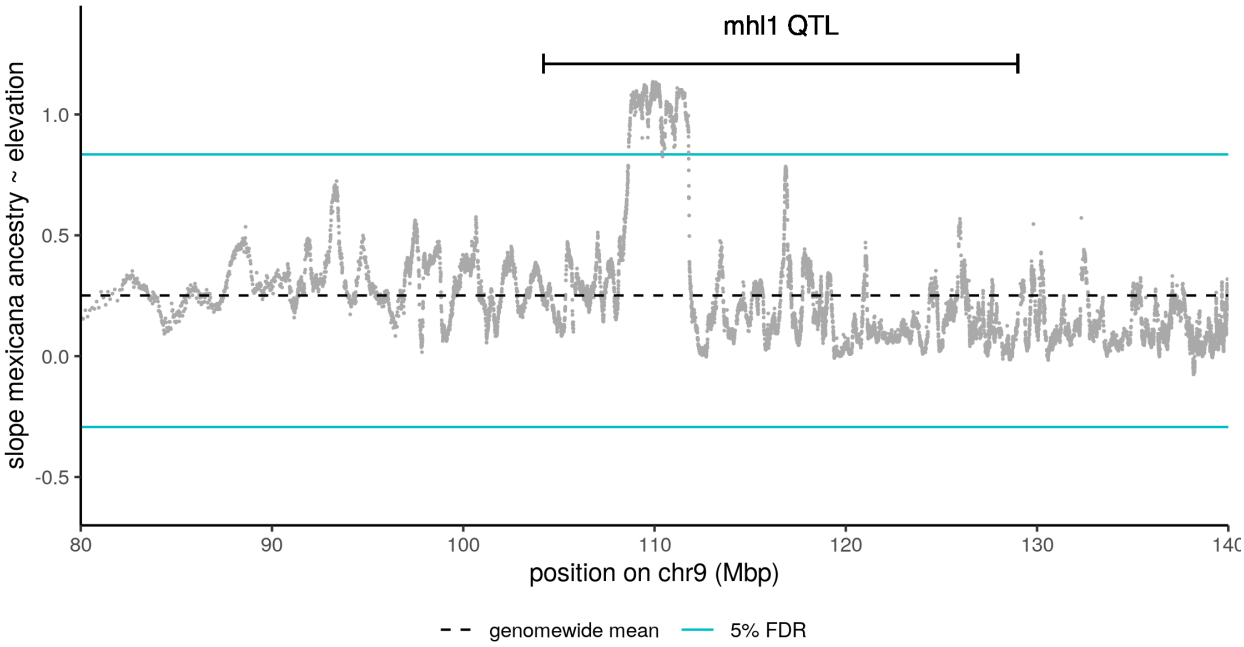


FIGURE 2.44. **Ancestry slope with elevation at mhl1 locus**. Slope of introgressed *mexicana* ancestry proportion in sympatric maize over a 1 km gain in elevation, zoomed in on the mhl1 QTL region on chromosome 9. Coordinates for the 3 Mb outlier region within this QTL are 9:108640415-111788150.

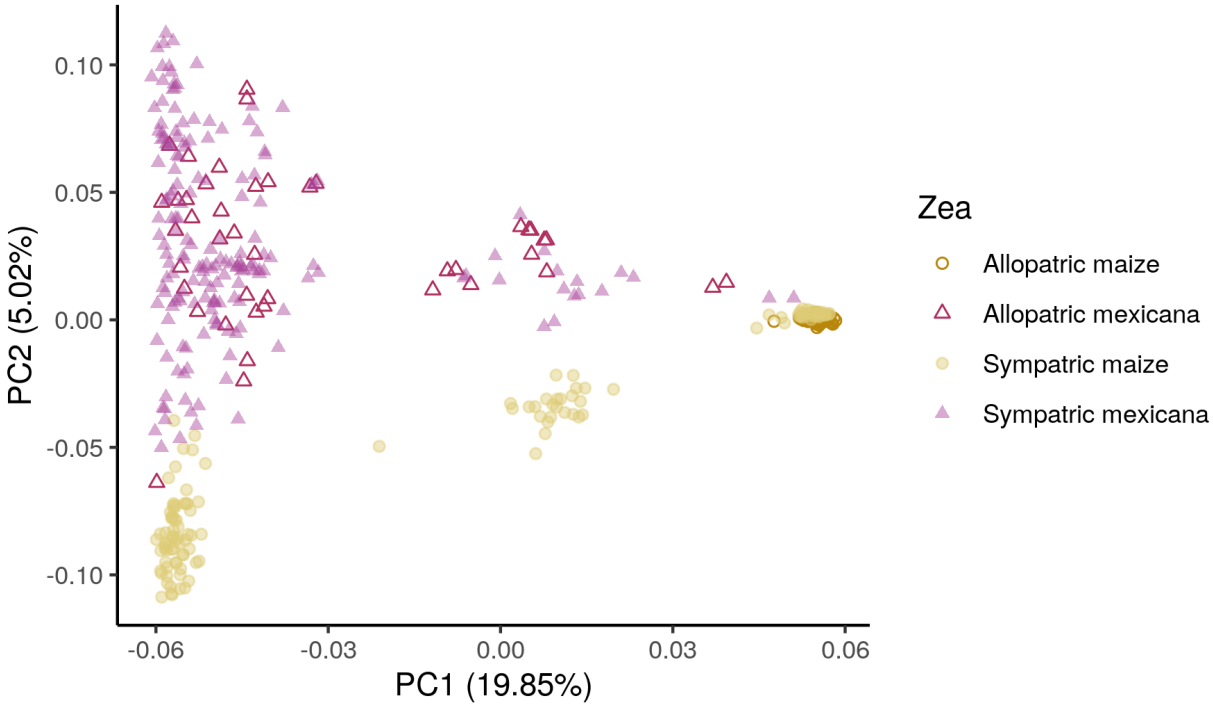


FIGURE 2.45. **PCA of putative mhl1 inversion**. Principal components analysis of all SNPs in the 3 Mb outlier region within the mhl1 QTL region that shows a steep increase in introgressed *mexicana* ancestry across elevation (>5% FDR). This region on chromosome 9 is a putative inversion (9:108640415-111788150), separating out into three clusters across PC1: individuals homozygous for the common *mexicana* inversion allele (left), heterozygous individuals (middle) and individuals homozygous for the common maize inversion allele (right; includes all allopatric reference maize).

Bibliography

[1] Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE, et al. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Research. 2008;18(1):67–76.

[2] Tavares H, Whibley A, Field DL, Bradley D, Couchman M, Copsey L, et al. Selection and gene flow shape genomic islands that control floral guides. Proceedings of the National Academy of Sciences USA. 2018;5:201801832.

[3] Powell DL, García-Olazábal M, Keegan M, Reilly P, Du K, Díaz-Loyo AP, et al. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. Science. 2020;368(6492):731–736.

[4] Hodgson JA, Pickrell JK, Pearson LN, Quillen EE, Prista A, Rocha J, et al. Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. Proceedings Biological sciences. 2014;281:20140930.

[5] Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. The genomic signature of crop-wild introgression in maize. PLoS Genetics. 2013;9(5):e1003477.

[6] Fitzpatrick BM, Johnson JR, Kump DK, Smith JJ, Voss SR, Shaffer HB. Rapid spread of invasive genes into a threatened native species. Proceedings of the National Academy of Sciences USA. 2010;107(8):3606–3610.

[7] Bay RA, Taylor EB, Schluter D. Parallel introgression and selection on introduced alleles in a native species. Molecular Ecology. 2019;28(11):2802–2813.

[8] Cridland JM, Tsutsui ND, Ramírez SR. The complex demographic history and evolutionary origin of the western honey bee, *Apis mellifera*. Genome Biology and Evolution. 2017;9(2):457–472.

[9] Moritz RFA, Härtel S, Neumann P. Global invasions of the western honeybee (Apis mellifera) and the consequences for biodiversity. Écoscience. 2016;12(3):289–301.

[10] Crane E. The world history of beekeeping and honey hunting; 1999.

[11] Levine JM. Biological invasions. Current biology : CB. 2008;18(2):R57–60.

[12] Winston ML. The biology and management of Africanized honey bees. Annual Review of Entomology. 1992;37:173–193.

[13] Stort AC. Genetic Study of Aggressiveness of two Subspecies of Apis Mellifera in Brazil 1. Some Tests to Measure Aggressiveness. Journal of Apicultural Research. 1974;13(1):33–38.

[14] Collins AM, Rinderer TE, Harbo JR, Bolten AB. Colony Defense by Africanized and European Honey Bees. Science. 1982;218(4567):72–74.

[15] Hunt GJ, Guzman-Novoa E, Fondrk MK, Page RE. Quantitative trait loci for honey bee stinging behavior and body size. Genetics. 1998;148(3):1203–1213.

[16] Winston ML. Killer bees. The Africanized honey bee in the Americas. Cambridge, MA: Harvard University Press; 1992.

[17] Roell A, Whitehead H, Van Wyk J. Why the term Africanized bees is problematic in a racist society; 2020. Available from: `https://doi.org/10.6084/m9.figshare.12735452.v1`.

[18] Tsing AL. Empowering nature, or: some gleanings in bee culture. In: Yanagisako S, Delaney C, editors. Naturalizing Power. New York, NY: Routledge; 1995. p. 113–143.

[19] Ksiazek P. Africanized honey bees; 2007. Press release, Zak Gallery.

[20] Schumacher MJ, Egen NB. Significance of Africanized Bees for Public Health: A Review. Archives of Internal Medicine. 1995;155(19):2038–2043.

[21] Woyke J. Experiences with Apis mellifera adansonii in Brazil and in Poland. Apiacta. 1973;.

[22] Villa JD, Koeniger N, Rinderer TE. Overwintering of Africanized, European, and hybrid honey bees in Germany. Environmental Entomology. 1991;20(1):39–43.

[23] Taylor Jr OR, Spivak M. Climatic limits of tropical African honeybees in the Americas. Bee World. 1984;65(1):38–47.

[24] Harrison JF, Fewell JH, Anderson KE, Loper GM. Environmental physiology of the invasion of the Americas by Africanized honeybees. Integrative and Comparative Biology. 2006;46(6):1110–1122.

[25] Southwick EE, Roubik DW, Williams JM. Comparative energy balance in groups of Africanized and European honey bees: ecological implications. Comparative Biochemistry and Physiology. 1990;97(1):1–7.

[26] Sheppard WS, Rinderer TE, Mazzoli JA, Stelzer JA, Shimanuki H. Gene flow between African- and European-derived honey bee populations in Argentina. Nature. 1991;349(6312):782–784.

[27] Agra MN, Conte CA, Corva PM, Cladera JL, Lanzavecchia SB, Palacio MA. Molecular characterization of Apis mellifera colonies from Argentina: genotypic admixture associated with ecoclimatic regions and apicultural activities. Entomologia Experimentalis et Applicata. 2018;166(9):724–738.

[28] Pinto MA, Rubink WL, Patton JC, Coulson RN, Johnston JS. Africanization in the United States: replacement of feral European honeybees (*Apis mellifera* L.) by an African hybrid swarm. Genetics. 2005;170(4):1653–1665.

[29] Loper GM, Fewell J, Smith DR, Sheppard WS, Schiff N. Changes in the genetics of a population of feral honey bees (Apis mellifera L.) in S. Arizona after the impact of tracheal mites (Acarapis woodi), Varroa mites (Varroa jacobsoni) and Africanization. In: Hoopingarner R, Connor L, editors. Apiculture for the 21st Century. Cheshire, CT: Wicwas; 1999. p. 47–51.

[30] Kono Y, Kohn JR. Range and frequency of Africanized honey bees in California (USA). PLoS ONE. 2015;10(9):e0137407.

[31] Lin W, McBroome J, Rehman M, Johnson BR. Africanized bees extend their distribution in California. PLoS ONE. 2018;13(1):e0190604.

[32] Kadri SM, Harpur BA, Orsi RO, Zayed A. A variant reference data set for the Africanized honeybee, Apis mellifera. Scientific Data. 2016;3:160097.

[33] Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, et al. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. Nature Genetics. 2014;46(10):1081–1088.

[34] Cridland JM, Ramírez SR, Dean CA, Sciligo A, Tsutsui ND. Genome sequencing of museum specimens reveals rapid changes in the genetic composition of honey bees in California. Genome Biology and Evolution. 2018;10(2):458–472.

[35] Whitfield CW, Behura SK, Berlocher SH, Clark AG, Johnston JS, Sheppard WS, et al. Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. Science. 2006;314(5799):642–645.

[36] Bozek K, Rangel J, Arora J, Tin M, Crotteau E, Loper G, et al. Parallel genomic evolution of parasite tolerance in wild honey bee populations. bioRxiv. 2018;doi:10.1101/498436.

[37] Nelson RM, Wallberg A, Simões ZLP, Lawson DJ, Webster MT. Genome-wide analysis of admixture and adaptation in the Africanized honeybee. Molecular Ecology. 2017;26:3603–3617.

[38] Ruttner F. Honeybees of Tropical Africa. In: Biogeography and Taxonomy of Honeybees. Berlin: Springer; 1988. p. 199–227.

[39] Schneider SS, DeGrandi-Hoffman G, Smith DR. The African honey bee: Factors contributing to a successful biological invasion. Annual Review Entomology. 2004;49(1):351–376.

[40] Danka RG, Rinderer TE, Hellmich RL, Collins AM. Comparative toxicities of four topically applied insecticides to Africanized and European honey bees (Hymenoptera: Apidae). Journal of Economic Entomology. 1986;79(1):18–21.

[41] Guzman-Novoa E, Vandame R, Arechavaleta ME. Susceptibility of European and Africanized honey bees (Apis mellifera L.) to Varroa jacobsoni Oud. in Mexico. Apidologie. 1999;30(2-3):173–182.

[42] Vandame R, Morand S, Colin ME, Belzunces LP. Parasitism in the social bee Apis mellifera: quantifying costs and benefits of behavioral resistance to Varroa destructor mites. Apidologie. 2002;33(5):433–445.

[43] Guerra J, Goncalves LS, De Jong D. Africanized honey bees (Apis mellifera L.) are more efficient at removing worker brood artificially infested with the parasitic mite Varroa jacobsoni Oudemans than are Italian bees or Italian/Africanized hybrids. Genetics and Molecular Biology. 2000;23(1):89–92.

[44] Moretto G, de Mello LJ. Varroa jacobsoni infestation of adult Africanized and Italian honey bees (Apis mellifera) in mixed colonies in Brazil. Genetics and Molecular Biology. 1999;22(3):321–323.

[45] Medina-Flores CA, Guzman-Novoa E, Hamiduzzaman MM, Aréchiga-Flores CF, López-Carlos MA. Africanized honey bees (*Apis mellifera*) have low infestation levels of the mite Varroa destructor in different ecological regions in Mexico. Genetics and Molecular Research. 2014;13(3):7282–7293.

[46] Daly HV, Balling SS. Identification of Africanized honeybees in the Western Hemisphere by discriminant analysis. Journal of the Kansas Entomological Society. 1978;.

[47] Danka RG, Hellmich RL, Rinderer TE, Collins AM. Diet-selection ecology of tropically and temperately adapted honey-bees. Animal Behaviour. 1987;35(6):1858–1863.

[48] Fewell JH, Bertram SM. Evidence for genetic variation in worker task performance by African and European honey bees. Behavioral Ecology and Sociobiology. 2002;52(4):318–325.

[49] Rivera-Marchand B, Oskay D, Giray T. Gentle Africanized bees on an oceanic island. Evolutionary applications. 2012;5(7):746–756.

[50] Avalos A, Pan H, Li C, Acevedo-Gonzalez JP, Rendon G, Fields CJ, et al. A soft selective sweep during rapid evolution of gentle behaviour in an Africanized honeybee. Nature Communications. 2017;8(1):351.

[51] Winston ML, Otis GW, Taylor Jr OR. Absconding Behaviour of the Africanized Honeybee in South America. Journal of Apicultural Research. 1979;18(2):85–94.

[52] Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics. 2013;195(3):693–702.

[53] Corbett-Detig R, Nielsen R. A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy. PLoS Genetics. 2017;13(1):e1006529.

[54] Kent RB. The introduction and diffusion of the African honeybee in South America. Yearbook of the Association of Pacific Coast Geographers. 1988;50(1):21–43.

[55] USDA Agricultural Research Service. Spread of Africanized honey bees by year, by county; 2009. Available from: `https://www.ars.usda.gov/ARSUserFiles/20220500/New%20Bee%20Map09%20compressed.jpg`.

[56] Becker R, Wilks A. Constructing a Geographical Database. AT&T Bell Laboratories Statistics Research Report. 1995;95.2.

[57] R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: `https://www.R-project.org/`.

[58] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686. doi:10.21105/joss.01686.

[59] Abrahamovich AH, Atela O, De la Rúa P, Galián J. Assessment of the mitochondrial origin of honey bees from Argentina. Journal of Apicultural Research. 2015;46(3):191–194.

[60] Simmons AD, Thomas CD. Changes in Dispersal during Species' Range Expansions. American Naturalist. 2015;164(3):378–395.

[61] Cwynar LC, MacDonald GM. Geographical Variation of Lodgepole Pine in Relation to Population History. American Naturalist. 1987;129(3):463–469.

[62] Phillips BL, Brown GP, Webb JK, Shine R. Invasion and the evolution of speed in toads. Nature. 2006;439(7078):803–803.

[63] Hill JK, Thomas CD, Blakeley DS. Evolution of flight morphology in a butterfly that has recently expanded its geographic range. Oecologia. 1999;121(2):165–170.

[64] Daly HV, Hoelmer K, Gambino P. Clinal geographic variation in feral honey bees in California, USA. Apidologie. 1991;22(6):591–609.

[65] Wang S, Rohwer S, Delmore K, Irwin DE. Cross-decades stability of an avian hybrid zone. Journal of Evolutionary Biology. 2019;32(11):1242–1251.

[66] Szymura JM, Barton NH. Genetic analysis of a hybrid zone between the fire-bellied toads, Bombina bombina and B. variegata, near Cracow in southern Poland. Evolution. 1986;40(6):1141.

[67] Szymura JM, Barton NH. The genetic structure of the hybrid zone between the fire-bellied toads Bombina bombina and B. variegata: Comparisons between transects and between loci. Evolution. 1991;45(2):237.

[68] Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology. 2005;25(15):1965–1978.

[69] Gompert Z, Mandeville EG, Buerkle CA. Analysis of Population Genomic Data from Hybrid Zones. Annual Review of Ecology, Evolution, and Systematics. 2017;48(1):207–229.

[70] Barton NH. Multilocus Clines. Evolution. 1983;37(3):454.

[71] Excoffier L, Foll M, Petit RJ. Genetic consequences of range expansions. Annual Review of Ecology, Evolution, and Systematics. 2009;40(1):481–501.

[72] Hunt GJ, Amdam GV, Schlipalius D, Emore C, Sardesai N, Williams CE, et al. Behavioral genomics of honeybee foraging and nest defense. Die Naturwissenschaften. 2007;94(4):247–267.

[73] Tsuruda JM, Harris JW, Bourgeois L, Danka RG, Hunt GJ. High-resolution linkage analyses to identify genes that influence *Varroa* sensitive hygiene behavior in honey bees. PLoS ONE. 2012;7(11):e48276.

[74] Oxley PR, Spivak M, Oldroyd BP. Six quantitative trait loci influence task thresholds for hygienic behaviour in honeybees (*Apis mellifera*). Molecular Ecology. 2010;19(7):1452–1461.

[75] Spötter A, Gupta P, Nuernberg G, Reinsch N, Bienefeld K. Development of a 44K SNP assay focussing on the analysis of a varroa-specific defence behaviour in honey bees (*Apis mellifera carnica*). Molecular Ecology Resources. 2012;12(2):323–332.

[76] Arechavaleta-Velasco ME, Alcala-Escamilla K, Robles-Rios C, Tsuruda JM, Hunt GJ. Fine-scale linkage mapping reveals a small set of candidate genes influencing honey bee grooming behavior in response to *Varroa* mites. PLoS ONE. 2012;7(11):e47269.

[77] McDonnell CM, Alaux C, Parrinello H, Desvignes JP, Crauser D, Durbesson E, et al. Ecto- and endoparasite induce similar chemical and brain neurogenomic responses in the honey bee (Apis mellifera). BMC Ecology. 2013;13(1):1–15.

[78] Surlis C, Carolan JC, Coffey M, Kavanagh K. Quantitative proteomics reveals divergent responses in Apis mellifera worker and drone pupae to parasitization by Varroa destructor. Journal of Insect Physiology. 2018;107:291–301.

[79] Buggs R. Empirical study of hybrid zone movement. Heredity. 2007;99(3):301–312.

[80] Taylor SA, Larson EL, Harrison RG. Hybrid zones: windows on climate change. Trends in Ecology & Evolution. 2015;30(7):398–406.

[81] Good TP, Ellis JC, Annett CA, Pierotti R. Bounded hybrid superiority in an avian hybrid zone: effects of mate, diet, and habitat choice. Evolution. 2000;54(5):1774–1783.

[82] De La Torre AR, Wang T, Jaquish B, Aitken SN. Adaptation and exogenous selection in a Picea glauca × Picea engelmannii hybrid zone: implications for forest management under climate change. New Phytologist. 2014;201(2):687–699.

[83] Adrion JR, Hahn MW, Cooper BS. Revisiting classic clines in Drosophila melanogaster in the age of genomics. Trends in genetics : TIG. 2015;31(8):434–444.

[84] Rinderer TE, Sylvester HA, Brown MA, Villa JD, Pesante D, Collins AM. Field and simplified techniques for identifying Africanized and European honey bees. Apidologie. 1986;17(1):13–48.

[85] Currat M, Ruedi M, Petit RJ, Excoffier L. The hidden side of invasions: massive introgression by local genes. Evolution. 2008;62(8):1908–1920.

[86] Barton N, Bengtsson BO. The barrier to genetic exchange between hybridising populations. Heredity. 1986;56:357–376.

[87] Harpur BA, Kadri SM, Orsi RO, Whitfield CW, Zayed A. Defense response in Brazilian honey bees (Apis mellifera scutellata x spp.) is underpinned by complex patterns of admixture. Genome Biology and Evolution. 2020;.

[88] Goulson D, Nicholls E, Botías C, Rotheray EL. Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. Science. 2015;347(6229):1255957–1255957.

[89] Tange O. GNU Parallel 2018. Ole Tange; 2018. Available from: `https://doi.org/10.5281/zenodo.1146014`.

[90] Rowan BA, Heavens D, Feuerborn TR, Tock AJ, Henderson IR, Weigel D. An ultra high-density Arabidopsis thaliana crossover map that refines the influences of structural variation and epigenetic features. Genetics. 2019;213(3):771–787.

[91] Harpur BA, Kent CF, Molodtsova D, Lebon JMD, Alqarni AS, Owayss AA, et al. Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. Proceedings of the National Academy of Sciences USA. 2014;111(7):2614–2619.

[92] Wallberg A, Bunikis I, Pettersson OV, Mosbech MB, Childers AK, Evans JD, et al. A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. BMC Genomics. 2019;20(1):275.

[93] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357–359.

[94] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078–2079.

[95] Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC bioinformatics. 2014;15(1):356.

[96] Jones JC, Wallberg A, Christmas MJ, Kapheim KM, Webster MT, Singh N. Extreme differences in recombination rate between the genomes of a solitary and a social bee. Molecular Biology and Evolution. 2019;36(10):2277–2291.

[97] Meisner J, Albrechtsen A. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics. 2018;210(2):719–731.

[98] Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating Kinship in Admixed Populations. The American Journal of Human Genetics. 2012;91(1):122–138.

[99] Long JC. The genetic structure of admixed populations. Genetics. 1991;127(2):417–428.

[100] Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002. Available from: `http://www.stats.ox.ac.uk/pub/MASS4`.

[101] Hong Y. poibin: The Poisson Binomial Distribution; 2019. Available from: `https://CRAN.R-project.org/package=poibin`.

[102] Hijmans RJ. geosphere: Spherical Trigonometry; 2019. Available from: `https://CRAN.R-project.org/package=geosphere`.

[103] Bickel PJ, Boley N, Brown JB, Huang H, Zhang NR. Subsampling methods for genomic inference. The Annals of Applied Statistics. 2010;4(4):1660–1697.

[104] Padfield D, Matheson G. nls.multstart: Robust Non-Linear Regression using AIC Scores; 2018. Available from: `https://CRAN.R-project.org/package=nls.multstart`.

[105] Ruttner F. Morphometric Analysis and Classification. Berlin: Springer; 1988.

[106] Grinde KE, Brown LA, Reiner AP, Thornton TA, Browning SR. Genome-wide significance thresholds for admixture mapping studies. American Journal of Human Genetics. 2019;104(3):454–465.

[107] Siegmund D, Yakir B. The Statistics of Gene Mapping. Springer Science & Business Media; 2007.

[108] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England). 2010;26(6):841–842.

[109] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009;4(1):44–57.

[110] Benjamini, Yoav, Hochberg, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995;57(1):289–300.

[111] Arechavaleta-Velasco ME, Hunt GJ, Emore C. Quantitative trait loci that influence the expression of guarding and stinging behaviors of individual honey bees. Behavior Genetics. 2003;33(3):357–364.

[112] Solignac M, Mougel F, Vautrin D, Monnerot M, Cornuet JM. A third-generation microsatellite-based linkage map of the honey bee, Apis mellifera, and its comparison with the sequence-based physical map. Genome biology. 2007;8(4):R66.

[113] Harpur B. Hunt honey bee markers; 2020. Dryad. Available from: `https://doi.org/10.5061/dryad.ns1rn8ppp`.

[114] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403–410.

[115] Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Research. 2013;23(9):1514–1521.

[116] Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature. 2006;444(7118):512–512.

[117] Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. PLoS ONE. 2011;6(1):e16327.

[118] Powell DL, García-Olazábal M, Keegan M, Reilly P, Du K, Díaz-Loyo AP, et al. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. Science. 2020;368(6492):731–736. doi:10.1126/science.aba5216.

[119] Zuellig MP, Sweigart AL. Gene duplicates cause hybrid lethality between sympatric species of Mimulus. PLOS Genetics. 2018;14(4):e1007130. doi:10.1371/journal.pgen.1007130.

[120] Presgraves DC. The molecular evolutionary basis of species formation. Nature Reviews Genetics. 2010;11(33):175–180. doi:10.1038/nrg2718.

[121] Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and introgression between Mimulus nasutus and Mimulus guttatus. PLoS Genetics. 2014;10(6):e1004410.

[122] Aeschbacher S, Selby JP, Willis JH, Coop G. Population-genomic inference of the strength and timing of selection against gene flow. Proceedings of the National Academy of Sciences. 2017;114(27):7061–7066. doi:10.1073/pnas.1616755114.

[123] Kenney AM, Sweigart AL. Reproductive isolation and introgression between sympatric Mimulus species. Molecular Ecology. 2016;25(11):2499–2517.

[124] Nelson TC, Stathos AM, Vanderpool DD, Finseth FR, Yuan Yw, Fishman L. Ancient and recent introgression shape the evolutionary history of pollinator adaptation and speciation in a model monkeyflower radiation (Mimulus section Erythranthe). PLOS Genetics. 2021;17(2):e1009095. doi:10.1371/journal.pgen.1009095.

[125] Martin SH, Davey JW, Salazar C, Jiggins CD. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLOS Biology. 2019;17(2):e2006288. doi:10.1371/journal.pbio.2006288.

[126] Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, et al. Genomic architecture and introgression shape a butterfly radiation. Science. 2019;366(6465):594–599. doi:10.1126/science.aaw2090.

[127] Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507(74927492):354–357. doi:10.1038/nature12961.

[128] Harris K, Nielsen R. The Genetic Cost of Neanderthal Introgression. Genetics. 2016;203(2):881–891.

[129] Juric I, Aeschbacher S, Coop G. The Strength of Selection against Neanderthal Introgression. PLoS Genetics. 2016;12(11):e1006340.

[130] Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 2018;360(6389):656–660.

[131] Pfennig KS, Kelly AL, Pierce AA. Hybridization as a facilitator of species range expansion. Proceedings of the Royal Society B: Biological Sciences. 2016;283(1839):20161329. doi:10.1098/rspb.2016.1329.

[132] Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, et al. Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science (New York, NY). 2015;347(6217):1258524. doi:10.1126/science.1258524.

[133] Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, et al. Widespread introgression across a phylogeny of 155 Drosophila genomes. bioRxiv. 2021; p. 33. doi:https://doi.org/10.1101/2020.12.14.422758.

[134] Martin CH, Cutler JS, Friel JP, Dening Touokong C, Coop G, Wainwright PC. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. Evolution. 2015;69(6):1406–1422.

[135] Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, et al. Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. Science. 2003;301(5637):1211–1216. doi:10.1126/science.1086949.

[136] Pease JB, Haak DC, Hahn MW, Moyle LC. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLOS Biology. 2016;14(2):e1002379. doi:10.1371/journal.pbio.1002379.

[137] Eberlein C, Hénault M, Fijarczyk A, Charron G, Bouvier M, Kohn L, et al. Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. Nature Communications. 2019;10(1):923–923. doi:10.1038/s41467-019-08809-7.

[138] Tusso S, Nieuwenhuis BPS, Sedlazeck FJ, Davey JW, Jeffares DC, Wolf JBW. Ancestral Admixture Is the Main Determinant of Global Biodiversity in Fission Yeast. Molecular Biology and Evolution. 2019;36(9):1975–1989. doi:10.1093/molbev/msz126.

[139] Purugganan MD. Evolutionary Insights into the Nature of Plant Domestication. Current Biology. 2019;29(14):R705–R714. doi:10.1016/j.cub.2019.05.053.

[140] Janzen GM, Wang L, Hufford MB. The extent of adaptive wild introgression in crops. New Phytologist. 2019;221(3):1279–1288. doi:https://doi.org/10.1111/nph.15457.

[141] Ellstrand NC, Meirmans P, Rong J, Bartsch D, Ghosh A, de Jong TJ, et al. Introgression of Crop Alleles into Wild or Weedy Populations. Annual Review of Ecology, Evolution, and Systematics. 2013;44(1):325–345. doi:10.1146/annurev-ecolsys-110512-135840.

[142] Hake S, Ross-Ibarra J. Genetic, evolutionary and plant breeding insights from the domestication of maize. eLife. 2015;4:e05861. doi:10.7554/eLife.05861.

[143] Hufford MB, Bilinski P, Pyhäjärvi T, Ross-Ibarra J. Teosinte as a model system for population and ecological genomics. Trends in Genetics. 2012;28(12):606–615. doi:10.1016/j.tig.2012.08.004.

[144] Mammadov J, Buyyarapu R, Guttikonda SK, Parliament K, Abdurakhmonov IY, Kumpatla SP. Wild Relatives of Maize, Rice, Cotton, and Soybean: Treasure Troves for Tolerance to Biotic and Abiotic Stresses. Frontiers in Plant Science. 2018;9. doi:10.3389/fpls.2018.00886.

[145] Piperno DR, Ranere AJ, Holst I, Iriarte J, Dickau R. Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. Proceedings of the National Academy of Sciences. 2009;106(13):5019–5024. doi:10.1073/pnas.0812525106.

[146] González JdJS, Corral JAR, García GM, Ojeda GR, Larios LDlC, Holland JB, et al. Ecogeography of teosinte. PLOS ONE. 2018;13(2):e0192676. doi:10.1371/journal.pone.0192676.

[147] Ross-Ibarra J, Tenaillon M, Gaut BS. Historical Divergence and Gene Flow in the Genus Zea. Genetics. 2009;181(4):1399–1413. doi:10.1534/genetics.108.097238.

[148] Fustier MA, Martínez-Ainsworth NE, Aguirre-Liguori JA, Venon A, Corti H, Rousselet A, et al. Common gardens in teosintes reveal the establishment of a syndrome of adaptation to altitude. PLOS Genetics. 2019;15(12):e1008512. doi:10.1371/journal.pgen.1008512.

[149] Fustier MA, Brandenburg JT, Boitard S, Lapeyronnie J, Eguiarte LE, Vigouroux Y, et al. Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. Molecular Ecology. 2017;26(10):2738–2756. doi:https://doi.org/10.1111/mec.14082.

[150] Aguirre-Liguori JA, Gaut BS, Jaramillo-Correa JP, Tenaillon MI, Montes-Hernández S, García-Oliva F, et al. Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (Zea mays parviglumis and Zea mays mexicana). Molecular Ecology. 2019;28(11):2814–2830. doi:https://doi.org/10.1111/mec.15098.

[151] Rodríguez JG, Sánchez G, Baltazar BM, Cruz LLDl, Santacruz-Ruvalcaba F, Ron PJ, et al. Characterization of floral morphology and synchrony among Zea species in Mexico. Maydica. 2006; p. 383–398.

[152] Piazena H. The effect of altitude upon the solar UV-B and UV-A irradiance in the tropical Chilean Andes. Solar Energy. 1996;57(2):133–140. doi:10.1016/S0038-092X(96)00049-7.

[153] Piperno DR, Flannery KV. The earliest archaeological maize (Zea mays L.) from highland Mexico: New accelerator mass spectrometry dates and their implications. Proceedings of the National Academy of Sciences. 2001;98(4):2101–2103. doi:10.1073/pnas.98.4.2101.

[154] Wilkes HG. Teosinte: The Closest Relative of Maize. The Bussey Institute of Harvard University; 1967.

[155] Wilkes HG. Hybridization of maize and teosinte, in Mexico and Guatemala and improvement of maize. Economic Botany. 1977;31(3):254–293.

[156] Stapleton AE, Walbot V. Flavonoids Can Protect Maize DNA from the Induction of Ultraviolet Radiation Damage. Plant Physiology. 1994;105(3):881–889. doi:10.1104/pp.105.3.881.

[157] Barthakur N. Temperature differences between two pigmented types of corn plants. International Journal of Biometeorology. 1974;18(1):70–75. doi:10.1007/BF01450666.

[158] Moya-Raygoza G. Early Development of Leaf Trichomes Is Associated With Decreased Damage in Teosinte, Compared With Maize, by Spodoptera frugiperda (Lepidoptera: Noctuidae). Annals of the Entomological Society of America. 2016;109(5):737–743. doi:10.1093/aesa/saw049.

[159] Lauter N, Gustus C, Westerbergh A, Doebley J. The Inheritance and Evolution of Leaf Pigmentation and Pubescence in Teosinte. Genetics. 2004;167(4):1949–1959. doi:10.1534/genetics.104.026997.

[160] Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB. The interplay of demography and selection during maize domestication and expansion. Genome biology. 2017;18(1):215.

[161] Gonzalez-Segovia E, Pérez-Limon S, Cíntora-Martínez GC, Guerrero-Zavala A, Janzen GM, Hufford MB, et al. Characterization of introgression from the teosinte Zea mays ssp. mexicana to Mexican highland maize. PeerJ. 2019;7:e6815. doi:10.7717/peerj.6815.

[162] Meyer RS, Purugganan MD. Evolution of crop species: genetics of domestication and diversification. Nature Reviews Genetics. 2013;14(1212):840–852. doi:10.1038/nrg3605.

[163] Stitzer MC, Ross-Ibarra J. Maize domestication and gene interaction. New Phytologist. 2018;220(2):395–408. doi:https://doi.org/10.1111/nph.15350.

[164] Doebley J. The Genetics of Maize Evolution. Annual Review of Genetics. 2004;38(1):37–59. doi:10.1146/annurev.genet.38.072902.092425.

[165] Fukunaga K, Hill J, Vigouroux Y, Matsuoka Y, Sanchez G J, Liu K, et al. Genetic diversity and population structure of teosinte. Genetics. 2005;169(4):2241–2254.

[166] Pyhäjärvi T, Hufford MB, Mezmouk S, Ross-Ibarra J. Complex patterns of local adaptation in teosinte. Genome Biology and Evolution. 2013;5(9):1594–1609.

[167] O'Brien AM, Sawers RJH, Strauss SY, Ross-Ibarra J. Adaptive phenotypic divergence in an annual grass differs across biotic contexts. Evolution. 2019;73(11):2230–2246. doi:https://doi.org/10.1111/evo.13818.

[168] Staller JE. High Altitude Maize (Zea Mays L.) Cultivation and Endemism in the Lake Titicaca Basin. Journal of Botany Research. 2016;1(11). doi:High Altitude Maize (Zea Mays L.) Cultivation and Endemism in the Lake Titicaca Basin.

[169] Chen Q, Samayoa LF, Yang CJ, Bradbury PJ, Olukolu BA, Neumeyer MA, et al. The genetic architecture of the maize progenitor, teosinte, and how it was altered during maize domestication. PLOS Genetics. 2020;16(5):e1008791. doi:10.1371/journal.pgen.1008791.

[170] Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. PLOS Genetics. 2011;7(4):e1001373. doi:10.1371/journal.pgen.1001373.

[171] Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013;193(4):1233–1254.

[172] Veller C, Edelman NB, Muralidhar P, Nowak MA. Recombination, variance in genetic relatedness, and selection against introgressed DNA. bioRxiv. 2019; p. 846147.

[173] Ogut F, Bian Y, Bradbury PJ, Holland JB. Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. Heredity. 2015;114(6):552–563.

[174] Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science. 2009;326(5956):1112–1115. doi:10.1126/science.1178534.

[175] Le Corre V, Siol M, Vigouroux Y, Tenaillon MI, Délye C. Adaptive introgression from maize has facilitated the establishment of teosinte as a noxious weed in Europe. Proceedings of the National Academy of Sciences. 2020;117(41):25618–25627. doi:10.1073/pnas.2006633117.

[176] Moose SP, Lauter N, Carlson SR. The Maize macrohairless1 Locus Specifically Promotes Leaf Blade Macrohair Initiation and Responds to Factors Regulating Leaf Identity. Genetics. 2004;166(3):1451–1461. doi:10.1534/genetics.166.3.1451.

[177] Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546(76597659):524–527. doi:10.1038/nature22971.

[178] Crow T, Ta J, Nojoomi S, Aguilar-Rangel MR, Rodríguez JVT, Gates D, et al. Gene regulatory effects of a large chromosomal inversion in highland maize. PLOS Genetics. 2020;16(12):e1009213. doi:10.1371/journal.pgen.1009213.

[179] Doebley J, Stec A, Gustus C. Teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics. 1995;141(1):333–346.

[180] Doebley J, Stec A, Hubbard L. The evolution of apical dominance in maize. Nature. 1997;386(66246624):485–488. doi:10.1038/386485a0.

[181] Dong Z, Xiao Y, Govindarajulu R, Feil R, Siddoway ML, Nielsen T, et al. The regulatory landscape of a core maize domestication module controlling bud dormancy and growth repression. Nature Communications. 2019;10(11):3810. doi:10.1038/s41467-019-11774-w.

[182] Dorweiler J, Stec A, Kermicle J, Doebley J. Teosinte glume architecture 1: A Genetic Locus Controlling a Key Step in Maize Evolution. Science. 1993;262(5131):233–235. doi:10.1126/science.262.5131.233.

[183] Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, et al. The origin of the naked grains of maize. Nature. 2005;436(70517051):714–719. doi:10.1038/nature03863.

[184] Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES. Genetic diversity and selection in the maize starch pathway. Proceedings of the National Academy of Sciences. 2002;99(20):12959–12962. doi:10.1073/pnas.202476999.

[185] Wills DM, Fang Z, York AM, Holland JB, Doebley JF. Defining the Role of the MADS-Box Gene, Zea Agamous-like1, a Target of Selection During Maize Domestication. Journal of Heredity. 2018;109(3):333–338. doi:10.1093/jhered/esx073.

[186] Wills DM, Whipple CJ, Takuno S, Kursel LE, Shannon LM, Ross-Ibarra J, et al. From Many, One: Genetic Control of Prolificacy during Maize Domestication. PLoS Genetics. 2013;9(6). doi:10.1371/journal.pgen.1003604.

[187] Sosso D, Luo D, Li QB, Sasse J, Yang J, Gendrot G, et al. Seed filling in domesticated maize and rice depends on SWEET-mediated hexose transport. Nature Genetics. 2015;47(1212):1489–1493. doi:10.1038/ng.3422.

[188] Vollbrecht E, Springer PS, Goh L, Buckler Iv ES, Martienssen R. Architecture of floral branch systems in maize and related grasses. Nature. 2005;436(70547054):1119–1126. doi:10.1038/nature03892.

[189] Sigmon B, Vollbrecht E. Evidence of selection at the ramosa1 locus during maize domestication. Molecular Ecology. 2010;19(7):1296–1311. doi:https://doi.org/10.1111/j.1365-294X.2010.04562.x.

[190] Lin Z, Li X, Shannon LM, Yeh CT, Wang ML, Bai G, et al. Parallel domestication of the Shattering1 genes in cereals. Nature Genetics. 2012;44(6):720–724. doi:10.1038/ng.2281.

[191] Trimble L, Shuler S, Tracy WF. Characterization of Five Naturally Occurring Alleles at the Sugary1 Locus for Seed Composition, Seedling Emergence, and Isoamylase1 Activity. Crop Science. 2016;56(4):1927–1939. doi:https://doi.org/10.2135/cropsci2015.02.0117.

[192] Doebley J, Stec A. Genetic Analysis of the Morphological Differences between Maize and Teosinte. Genetics. 1991;129(1):285–295.

[193] Doebley J, Stec A. Inheritance of the morphological differences between maize and teosinte: comparison of results for two F2 populations. Genetics. 1993;134(2):559–570. doi:10.1093/genetics/134.2.559.

[194] Bomblies K, Doebley JF. Pleiotropic Effects of the Duplicate Maize FLORICAULA/LEAFY Genes zfl1 and zfl2 on Traits Under Selection During Maize Domestication. Genetics. 2006;172(1):519–531. doi:10.1534/genetics.105.048595.

[195] Gallavotti A, Zhao Q, Kyozuka J, Meeley RB, Ritter MK, Doebley JF, et al. The role of barren stalk1 in the architecture of maize. Nature. 2004;432(70177017):630–635. doi:10.1038/nature03148.

[196] Wang Z, Ueda T, Messing J. Characterization of the maize prolamin box-binding factor-1 (PBF-1) and its role in the developmental regulation of the zein multigene family. Gene. 1998;223(1):321–332. doi:10.1016/S0378-1119(98)00244-3.

[197] Rodríguez-Zapata F, Barnes AC, Blöcher-Juárez KA, Gates D, Kur A, Wang L, et al. Teosinte introgression modulates phosphatidylcholine levels and induces early maize flowering time. bioRxiv. 2021;doi:10.1101/2021.01.25.426574.

[198] Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M. A Gene Regulatory Network Model for Floral Transition of the Shoot Apex in Maize and Its Dynamic Modeling. PLOS ONE. 2012;7(8):e43450. doi:10.1371/journal.pone.0043450.

[199] Li Yx, Li C, Bradbury PJ, Liu X, Lu F, Romay CM, et al. Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. The Plant Journal. 2016;86(5):391–402. doi:https://doi.org/10.1111/tpj.13174.

[200] Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, et al. The genetic architecture of maize flowering time. Science. 2009;325(5941):714–718. doi:10.1126/science.1174276.

[201] R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: `https://www.R-project.org/`.

[202] Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research. 2018;7:1338. doi:10.12688/f1000research.15931.2.

[203] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

[204] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio]. 2013;.

[205] Li H. Improving SNP discovery by base alignment quality. Bioinformatics. 2011;27(8):1157–1158. doi:10.1093/bioinformatics/btr076.

[206] Yang CJ, Samayoa LF, Bradbury PJ, Olukolu BA, Xue W, York AM, et al. The genetic architecture of teosinte catalyzed and constrained maize domestication. Proceedings of the National Academy of Sciences. 2019;116(12):5643–5652. doi:10.1073/pnas.1820997116.

[207] van Heerwaarden J, Ross-Ibarra J, Doebley J, Glaubitz JC, Gonzalez JdJS, Gaut BS, et al. Fine scale genetic structure in the wild ancestor of maize (Zea mays ssp. parviglumis). Molecular ecology. 2010;doi:10.1111/j.1365-294X.2010.04559.x.

[208] Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proceedings of the National Academy of Sciences. 2001;98(16):9161–9166. doi:10.1073/pnas.151244298.

[209] Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;132(2):583–589.

[210] Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: The impact of rare variants. Genome Research. 2013;23(9):1514–1521. doi:10.1101/gr.154831.113.

[211] Haider S, Waggott D, C Boutros P. bedr: Genomic Region Processing using Tools Such as 'BEDTools', 'BEDOPS' and 'Tabix'; 2019. Available from: `https://CRAN.R-project.org/package=bedr`.

[212] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328(5979):710–722. doi:10.1126/science.1188021.

[213] Peter BM. Admixture, Population Structure, and F-Statistics. Genetics. 2016;202(4):1485–1501. doi:10.1534/genetics.115.183913.

[214] Soraggi S, Wiuf C, Albrechtsen A. Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. G3: Genes—Genomes—Genetics. 2018;8(2):551–566. doi:10.1534/g3.117.300192.

[215] Portwood I John L, Woodhouse MR, Cannon EK, Gardiner JM, Harper LC, Schaeffer ML, et al. MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Research. 2019;47(D1):D1146–D1154. doi:10.1093/nar/gky1046.

[216] Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2522. doi:10.1093/bioinformatics/bts480.

[217] Garnier S. viridis: Default Color Maps from 'matplotlib'; 2018. Available from: `https://CRAN.R-project.org/package=viridis`.

[218] Canty A, Ripley BD. boot: Bootstrap R (S-Plus) Functions; 2019.

[219] Davison AC, Hinkley DV. Bootstrap Methods and Their Applications. Cambridge: Cambridge University Press; 1997. Available from: `http://statwww.epfl.ch/davison/BMA/`.

[220] Auguie B. gridExtra: Miscellaneous Functions for "Grid" Graphics; 2017. Available from: `https://CRAN.R-project.org/package=gridExtra`.

[221] Ahlmann-Eltze C. ggupset: Combination Matrix Axis for 'ggplot2' to Create 'UpSet' Plots; 2020. Available from: `https://CRAN.R-project.org/package=ggupset`.

[222] Pedersen TL. tidygraph: A Tidy API for Graph Manipulation; 2020. Available from: `https://CRAN.R-project.org/package=tidygraph`.