Environment but not geography explains genetic variation in the invasive and largely
panmictic European starling in North America

Running title: Genotype-environment associations in starlings

Natalie R. Hofmeister^{1,2*}, Scott J. Werner³, and Irby J. Lovette^{1,2}

¹ Department of Ecology and Evolutionary Biology, Cornell University, 215 Tower Road,
Ithaca, NY 14853

² Fuller Evolutionary Biology Program, Cornell Lab of Ornithology, Cornell University, 159
Sapsucker Woods Road, Ithaca, NY 14850

³ United States Department of Agriculture, Animal and Plant Health Inspection Service,
Wildlife Services, National Wildlife Research Center, 4101 LaPorte Avenue, Fort Collins, CO
80521

*corresponding author: nrh44@cornell.edu

ABSTRACT

Populations of invasive species that colonize and spread in novel environments may
differentiate both through demographic processes and local selection throughout the
genome. European starlings (*Sturnus vulgaris*) were introduced to New York in 1890 and
subsequently spread throughout North America, becoming one of the most widespread and
numerous bird species on the continent. Genome-wide comparisons across starling
individuals and populations can identify demographic and/or selective factors that

facilitated this rapid and successful expansion. We investigated patterns of genomic diversity and differentiation using reduced-representation genome sequencing (ddRADseq) of 17 starling populations. Consistent with this species' high dispersal rates and rapid expansion history, we found low genome-wide differentiation and few F_{ST} outliers even at a continental scale. Despite starting from a founding population of approximately 180 individuals, North American starlings do not seem to have undergone a detectable genetic bottleneck: they have maintained an extremely large effective population size since introduction. We find more than 200 variants that correlate with temperature and/or precipitation. Genotype-environment associations (but not outlier scans) identify these SNPs against a background of negligible genome- and range-wide divergence. Such variants fall in the coding regions of genes associated with metabolism, stress, and neurological function. This evidence for incipient local adaptation in North American starlings suggests that it can evolve rapidly even in wide-ranging and evolutionarily young populations. This survey of genomic signatures of expansion in North American starlings is the most comprehensive to date and complements ongoing studies of world-wide local adaptation in these highly dispersive and invasive birds.

Keywords: panmixia, gene flow, local adaptation, invasion genetics

INTRODUCTION

Studies of local adaptation have long bridged the interface between ecological and evolutionary questions by exploring how populations adapt to differing environmental conditions. Traditionally, high degrees of local adaptation were expected to be present only

in fairly isolated populations—those free from the homogenizing effects of high gene flow—with a long history in those locations, providing the time thought to be necessary for local adaptation to evolve (Lenormand 2002). In contrast, we now know that local adaptation occurs frequently even in systems with high gene flow (Yeaman & Whitlock 2011; Tigano & Friesen 2016) and often rapidly after colonization of a novel environment (Prentis et al. 2008). We continue to find evidence for rapid local adaptation in systems as divergent as cane toads (Rollins et al. 2015), sticklebacks (Lescak et al. 2015), honeybees (Avalos et al. 2017), steelhead trout (Willoughby *et al.* 2018), deer mice (Pfeifer *et al.* 2018), and many more. These studies show that many taxa can adapt rapidly to local conditions in response to the new selection regimes they encounter as they expand their range. Invasive species that have recently expanded into new locations provide tractable opportunities to investigate local adaptation as it originates (Colautti & Lau 2015).

After successful colonization of a new habitat, many invasive species show a demographic boom that is likely facilitated by their ecological release in the new environment. Theory predicts that this rapid population growth will plateau as the population approaches carrying capacity, but successful invasive species may continue to expand their range and thus maintain a high rate of overall population growth. When population density increases and demographic rates change, introduced species may rapidly evolve traits that enable them to spread (Szűcs et al. 2017). Increased dispersal may evolve in concert with the demographic boom, enabling populations to grow and spread as individuals disperse. This dispersal can result in gene flow that counteracts inbreeding depression and increases adaptive potential (Garant *et al.* 2007; Rius & Darling 2014). If particular traits enable individuals to disperse more easily to their preferred

habitat, gene flow may be directional and even adaptive (Edelaar & Bolnick 2012; Jacob *et al.* 2017). Dispersal with habitat choice has thus facilitated range expansion in Western bluebirds (Duckworth 2008) and invasive beetles (Lombaert *et al.* 2014; Ochocki & Miller 2017) among other species. For example, invasive European starlings in South Africa disperse more frequently at the leading edge of their range expansion, where rates of dispersal are determined by demographic changes and environmental quality (Hui *et al.* 2012). Invasions thus allow us to observe interactions between demography and the early processes of selection (Dlugosch *et al.* 2015) as populations experience new environments. Importantly, these eco-evolutionary interactions depend in part on the genetic variation in the established population.

Recent work in invasion genetics aims to tease apart how selection and demography might resolve a paradox of invasion (Estoup *et al.* 2016): many invasive species experience genetic bottlenecks as a result of an initial founder effect, but often thrive and spread despite this loss of standing genetic diversity. Theory predicts that introductions will typically result in an initial contraction in population size and/or genetic diversity (Dlugosch & Parker 2008). However, bottlenecks of genetic variation clearly do not limit the success of many invasive species (Schmid-Hempel *et al.* 2007; Dlugosch & Parker 2008; Facon *et al.* 2011). Invaders might adapt through soft sweeps that reduce genetic diversity while selecting for adaptive variants from standing or novel genetic variation, which is especially likely in the case of ecological adaptation (Messer & Petrov 2013). Furthermore, some invasions may increase rather than reduce genetic diversity, as when multiple invasions from different source populations introduce previously isolated alleles and thereby facilitate admixture (Dlugosch & Parker 2008). The new conditions can also select

among standing variation, where the presence of certain genetic variants in the native range accelerates adaptation upon introduction (Tsutsui *et al.* 2000; Schlaepfer *et al.* 2009; Hufbauer *et al.* 2011). In sum, although genetic diversity in introduced populations is often viewed as a pre-requisite to adaptation, changes in genetic diversity alone do not explain invasiveness (Uller & Leimu 2011).

The European starling (*Sturnus vulgaris*) stands out as an exceptionally successful avian colonist and invasive species. In North America, an estimated 200 million starlings currently range from northern Mexico to southern Alaska. Introduced to New York City in 1890, starlings nearly covered the continent within a few generations by expanding up to 91 km each year (Bitton & Graham 2014). The starling population has grown steadily and earned its reputation as a costly agricultural pest, as flocks often overwinter in feed lots or swarm fruit orchards (Linz *et al.* 2007). Starlings also may compete for nest sites with American Kestrels (*Falco sparverius*), Eastern Bluebirds (*Sialia sialis*), and other native cavity-nesting birds, though this competition may not have substantial demographic consequences for most native birds (Koenig 2003). Despite the starling's prominence as the most successful avian invasive species in North America, not much is known about the genetic consequences of its invasion history.

Genetic studies of other starling invasions suggest that North American starlings could have differentiated or adapted even after a few generations. Previous genetic work in North America indicated near-random mating at a continental scale, with large demes and high dispersal rates (Cabe 1998; 1999). However, these studies relied on a handful of microsatellite markers that sample a far smaller fraction of the genome than do current genomic tools. Additional studies in the invasive Australian population, which colonized

that continent nearly concurrently with the North American invasion, used multiple genetic approaches. Microsatellite evidence suggests that gene flow among Australian starling populations is low (Rollins *et al.* 2009), and phylogeographic patterns of mitochondrial sequence variation confirm that starlings on the edge of the expansion front in Western Australia have differentiated from those still living in the introduction site (Rollins *et al.* 2011). In fact, starlings at the expansion front may have rapidly adapted during the Australian invasion: the proportion of adult starlings in Western Australia carrying a novel mitochondrial haplotype has increased rapidly only at this range edge (Rollins *et al.* 2016). A genotyping-by-sequencing survey employing a much greater number of SNP markers indicates three populations in Australia, and geographic distance explains genetic differentiation in starlings better than does environmental variation in the Australian invasion (Cardilini 2016). Studies of the Australian invasion suggest incipient ecological specialization at the range edge; thus, our study of the North American starling invasion begins by examining the relationship between genetics and geography prior to testing for environmental associations (Sexton *et al.* 2014).

Here we explore the genomic and demographic patterns of range expansion in North American starlings with three specific aims: (1) to characterize genome-wide levels of diversity and differentiation among starlings; (2) to examine how genetic variation changes across the North American range; and (3) to test for signatures of selection associated with environmental gradients. Although traditional outlier-based methods may not recover evidence of local adaptation in a species that likely has low overall levels of genetic diversity, recent developments in genotype-environment association methods can identify polygenic traits which are locally adapted to overlapping environmental gradients

(Forester *et al.* 2018; Capblancq *et al.* 2018). Especially with these more sensitive methods, the same low genetic diversity resulting from a founder effect could improve our ability to discriminate signatures of selection from low background divergence across the genome (Dlugosch *et al.* 2015). This study thereby leverages modern genomic and analytical tools to examine the evolutionary history of this infamous avian invasion.

METHODS

Sample collection and processing

Starlings ($N = 266$) were collected in January-March 2016 and 2017 from 26 dairies and feedlots by the U.S. Department of Agriculture's Wildlife Services personnel in Arizona, California, Colorado, Idaho, Illinois, Iowa, Kansas, Missouri, Nebraska, Nevada, New Hampshire, New Mexico, New York, North Carolina, Texas, Washington, and Wisconsin. Whole birds were euthanized and stored at 0°C until tissue sampling. Breast muscle tissue was sampled using biopsy punches (Integra Miltex, York, PA) and frozen in 95% ethanol. Samples were shipped on dry ice, and DNA was extracted using a Qiagen DNeasy kit following the manufacturer protocol (Qiagen, New York, NY). DNA concentration of each sample was quantified using a Qubit fluorometer (Life Technologies, New York, NY). The collection and use of starlings for this and related studies were approved by the U.S. Department of Agriculture, National Wildlife Research Center's Institutional Animal Care and Use Committee (QA-2572, S.J. Werner- Study Director; Werner *et al.*, *in prep*; Supplementary Table 1).

Following the protocol of Peterson *et al.* 2012, we generated a reduced-representation genomic dataset of double-digested, restriction-site associated DNA

markers as described in Thrasher et al. 2017 using the restriction enzymes SbfI and MspI and adaptors P1 and P2. We trimmed and filtered for quality using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit). We then used the process_radtags commands in STACKS v 1.19 (Catchen et al. 2013) to demultiplex the remaining sequences. In subsequent filtering steps, we retained reads only if the following conditions were met: reads passed the Illumina chastity filter, contained an intact *SbfI* RAD site, contained one of the unique barcodes, and did not contain Illumina indexing adaptors. We also removed two individuals with >50% missing data and >50% relatedness (measured using the unadjusted AJK statistic and calculated within *vcftools*), leaving 158 individuals remaining in the sample.

We assembled sequences to an *S. vulgaris* reference genome (Hofmeister, Rollins et al., *in prep*) using the ref-map option in STACKS. Individual reads were mapped to the reference genome using BOWTIE2 version 2.2.8 (Langmead & Salzberg 2012) using the “very sensitive local” set of alignment pre-sets. The bioinformatics pipeline used for the reference-based assembly has the advantage of using less similarity thresholds to build loci. We required that a SNP be present in a minimum of 80% of the individuals with a minimum stack depth of 10 for individuals at a locus for it to be called. For analyses such as BayeScan we used all SNPs in a given stack, but for STRUCTURE and other analyses sensitive to linkage disequilibrium, we used only the first SNP in each stack.

(1) Patterns of genetic diversity and differentiation

We estimated per-locus measures of genetic diversity and genome-wide differentiation using the *populations* option in Stacks (Catchen et al. 2013). We used *vcftools* to calculate F_{ST} among population pairs and heterozygosity and nucleotide

diversity (π) within populations (Danecek et al. 2011). We investigated genetic structure within North American starlings using an analysis of molecular variance (AMOVA) in the R package *poppr* (Kamvar et al. 2014). We tested whether most genetic variation was observed among individuals or among sampling sites (“populations”). To determine significance, we compared observed variation at each hierarchical level to the randomly permuted distance matrices for that particular level using 1000 simulations in the function *randtest()* in the R package *adeigenet* (Jombart 2008), hypothesizing that the observed variance is greater than expected within individuals and less between individuals and between populations. We tested for isolation by distance (IBD) using a simple Mantel test in *adeigenet* (Jombart 2008): for these data, the assumption of stationarity likely holds, given that North American starlings appear to be in mutation-migration-drift equilibrium (Guillot & Rousset 2013).

(2) Population structure

We first tested for population structure using STRUCTURE (Pritchard *et al.* 2000) by simulating 10 runs of each K , hypothesizing that North American starlings would cluster into at most eight populations ($K=1-8$). To select the best-supported K , we used the Evanno method implemented in STRUCTURE HARVESTER v0.6.94 (Earl & vonHoldt 2011). We averaged results across the 10 runs using the greedy algorithm in the program CLUMPP v1.1.2 (Jakobsson & Rosenberg 2007), and visualized results using DISTRUCT v1.1 (Rosenberg 2003). Given that evidence of population structure depends on the filtering thresholds selected, we ran this model-based approach using a very strict minimum minor allele frequencies (MAF=0.3) and a more relaxed minimum frequency (MAF=0.1) (Linck &

Batthey 2017). STRUCTURE results did not differ substantially among MAF thresholds, and we used a filtered dataset with a minimum minor allele frequency of 0.1 in subsequent analyses. We also used non-parametric approaches to determine whether starlings clustered by sampling location, using principal components analysis in SNPRelate (Zheng et al. 2012) and discriminant analysis of principal components in adegenet (Jombart 2008).

Because we expect population structure to be fairly low given the recent expansion of North American starlings, we used fineRADstructure to test for more subtle patterns of structure (Malinsky et al. 2016). This program calculates shared ancestry using a coalescent model to determine haplotype linkage among sampled individuals. The resulting coancestry matrix controls for similarity among individuals to infer fine-scale patterns of population structure.

To identify potential geographic barriers, we used the program EEMS (Estimated Effective Migration Surfaces, (Petkova et al. 2015). EEMS estimates how quickly genetic similarity decays across the landscape, allowing us to pinpoint geographic regions that depart from continuous IBD. Because the number of hypothesized demes (subpopulations) can influence model sensitivity, we ran EEMS using polygons covering the entire North American range and only the areas sampled, and also tested each map using different numbers of demes (N=200, N=500, N=750 and N=1000). We adjusted the variance of the proposal distribution for both migration and diversity parameters to ensure that all parameters were accepted between 10 and 15% of the time, with the input proposal variances as follows: mSeedsProposalS2 = 0.15, mEffctProposalS2 = 1.2, qSeedsProposalS2 = 1.8, qEffctProposalS2 = 0.1, and mrateMuProposalS2 = 0.01.

(3) Demographic modeling

Because starlings appear to be panmictic based on evidence from STRUCTURE and PCA, we reconstructed demographic history using the Stairway plot method (Liu & Fu 2015). This method estimates recent population histories from hundreds of unphased, low-coverage loci, which distinguishes the stairway plot from other demographic methods (e.g., PSMC) that can infer ancient population history more accurately. The stairway plot method models changes in population size using the site frequency spectrum, where the null model assumes constant size. We used this model-flexible method to determine whether starlings experienced any genetic bottleneck after introduction: in the stairway plot, this result could occur if an alternative model was accepted during one or more steps of the stairway plot. We assumed a mutation rate of 1×10^{-9} and a generation length of 4.6 years (BirdLife International). We estimated the folded SFS using a Python script from Simon Martin (available at https://github.com/simonhmartin/genomics_general), and used the recommended 67% of sites for training. The results presented here are averaged among eight independent runs, each with 10 to 30 randomly generated breakpoints during the reconstruction.

(4) Selection analyses

We first identified environmental variation at each sampling location using the R package *raster* (Hijmans & van Etten 2012), extracting all 19 bioclimatic variables for each set of sampling coordinates from WorldClim 2.0 at a resolution of 5 min on June 16, 2018 (Fick & Hijmans 2017). For the univariate method (LFMM), environmental variation was modeled as the first three principal components of bioclimatic variation across the range of

North American starlings. The multivariate methods (RDA) used raw values of bioclimatic variables, retaining five variables with relatively low variance inflation factors: BIO1 (VIF=3.54), BIO7 (4.55), BIO12 (8.69), BIO16 (7.91), and elevation (2.19).

As a univariate test of selection, we used the *lmm* function (Frichot *et al.* 2013) to test for associations with climatic gradients and to decrease the number of false positives. We used the R package *LEA* (Frichot & François 2015) to prepare input files and run a model where genotypic variation is considered a response variable in a linear regression that controls for latent factors (e.g., population structure and/or background variation) in estimating the association between the genotypic response and the environmental predictor. For each of three models—including 1, 2, and 3 latent factors—we ran 30 MCMC chains of 10,000 cycles each, discarding a burn-in of 5,000 cycles. Z-scores were combined across all 30 runs and p-values readjusted to calibrate the null hypothesis and increase power using the Fisher-Stouffer method as suggested in the *LEA* and *LFMM* manuals. We used the Benjamini-Hochberg algorithm to control for false discoveries. Since we identified 1315 candidates using a q-value cut-off of 0.01—which is already 10-fold more stringent than the recommended value—only loci that were identified in all three runs (K=1-3 latent factors) and were more than five standard deviations from the mean log₁₀p value were considered candidates under selection (FDR<0.05, TP>25).

To complement the univariate methods (Bayescenv and LFMM), We also used redundancy analysis (RDA) to examine how loci may covary across multiple environmental gradients (Forester *et al.* 2018). RDA is especially powerful when testing for weak selection, and detects true positives in large data sets more reliably than other multivariate methods like Random Forest (Forester *et al.* 2018). Because RDA requires no missing data,

we first imputed genotypes where missing sites were assigned the genotype of highest probability across all individuals—a conservative but quick imputation method. Out of over 1.6 million data points, about 750,000 were missing. We then used the R package *vegan* to run the RDA model; for a full description of this method, see (Forester *et al.* 2018). Briefly, RDA uses constrained ordination to model a set of predictor variables, and unconstrained ordination axes to model the response (genetic variation). RDA infers selection on a particular locus when it loads heavily onto one or more unconstrained predictor axes; we retained five predictors (BIO1, BIO7, BIO12, BIO16, and elevation) with relatively low variance inflation factors (range: 2.2—8.7). We tested for significance using the *anova.cca* function within the *vegan* package, and also permuted predictor values across individuals to further check significance of the model. We identify candidate loci as those that are 3 or more standard deviations outside the mean loading. The R script for all RDA analyses and figures were written by Brenna Forester (available at https://popgen.nescent.org/2018-03-27_RDA_GEA.html).

Differentiation methods can identify loci that have undergone strong selective sweeps, but these methods may be inappropriate in systems like this one with low overall differentiation like this one. On the other hand, F_{ST} -based genome-scan approaches may identify loci that stand out against a low background level of differentiation. To test these more traditional methods against the model-based approaches described above, we also used BayeScEnv, which incorporates environmental differentiation when identifying outlier loci (de Villemereuil & Gaggiotti 2015). BayeScEnv includes a term to explicitly model environmental differentiation in the framework used in BayeScan, which enables us

to pinpoint loci that may be associated with environmental variation. For full details on this method, see the Supplementary Information.

To functionally annotate the contigs that contain these candidate SNPs, we assessed homology to the starling (*Sturnus vulgaris*) genome using BLASTN (Altschul et al. 1990). We first used the bedtools “getfasta” option to extract 10kb regions surrounding each candidate SNP (Quinlan & Hall 2010). We identified genes by choosing the match with the lowest E-value and highest identity: most matches showed 100% identity and an E-value approximated to zero, and all matches have an E value of $< 1.0 \times 10^{-5}$ and $> 90\%$ identity. Gene IDs of all candidates were uploaded to PANTHER to test for overrepresentation of particular gene ontology terms using a Fisher’s exact test and displaying the false discovery rate (Thomas *et al.* 2003; Mi *et al.* 2009).

RESULTS

(1) Patterns of genetic diversity and differentiation

We identified 15,038 SNPs at a mean of 27X coverage across 17 sampling locations of European starlings in North America. Genome-wide F_{ST} is extremely low (0.0085), and measures of genetic diversity do not vary substantially among sampling locations (Table 1). Across all populations, the highest pairwise F_{ST} was 0.0106, differentiating birds from the adjacent states of Arizona and New Mexico. Using a haplotype-based statistic of differentiation, ϕ_{ST} among populations shows an absence of genetic structure ($\phi_{ST} = 0.0002$). Hierarchical AMOVAs reveal that 94% of the observed genetic variance is explained by variation within individuals, and the remaining variance reflects differences among individuals in the same population, with negligible variation explained at the

between-population level (Figure 1C-D). Across the genome, F_{ST} and nucleotide diversity are exceptionally low (Figure 2A-B). Genome-wide heterozygosity is moderate at 0.339, and observed heterozygosity differs significantly from expected: there are more loci with a minor allele frequency near zero than might be expected by chance ($t = 66.6$, $df = 3569$, $P < 0.001$; Figure 2D), although the genetic mechanism generating this excess of rare alleles is unknown.

Figure 1. Population structure. A) Principal components analysis on 6,287 SNPs, explaining 1.03-1.08% of genetic variation observed. East and west indicate sampling locations relative to the Rocky Mountains. B) STRUCTURE analyses with K=2 and K=3 (best supported). C) Significance testing of hierarchical AMOVAs: the histogram shows expected variance components based on simulations, and the black diamond is the observed variance component. D) AMOVA results.

(2) Population structure

Principal components explain only 0.91% of variation among individuals (Figure 1A), and although STRUCTURE identified three populations at the best-supported value of K, these predicted populations do not show obvious differences in ancestry proportions (Figure 1B). Controlling for shared ancestry does not resolve population structure, and instead provides support for uniform gene flow among individuals (Figure S1). K-means clustering within DAPC also does not identify biologically relevant clusters.

There is no relationship between geographic and genetic distance (Mantel statistic = -0.07943, $P=0.904$). Spatially explicit models of isolation-by-distance show that starlings west of the Rocky Mountains are more diverse than their eastern counterparts (Figure 3). In terms of migration, the model correctly recovers a higher rate of migration of starlings moving north from the introduction site in New York, which matches what we know from historical records (Jernelov 2017). Although the EEMS model estimates the highest migration rate in the south of the North American range, we have few samples within the area of highest inferred migration. The model of migration estimates the lowest rates of migration among Midwest populations, where sampling is most thorough. The migration rate is extremely low just east of the potential geographic barrier of the Rocky Mountains, but increases strongly just west of the mountains.

Figure 2. Estimates of diversity and migration rates. Warmer (red) colors indicate lower levels of migration and diversity, whereas cool (blue) colors show high rates. The upper panel shows that diversity is estimated to be highest in the WA and NV populations, and diversity is much lower in the Midwest. The lower panel indicates low migration rates among nearly all populations sampled.

(3) Demographic modeling

For much of the starling's residence in North America, models suggest that its effective population size has declined steadily (Figure S2). Upon introduction approximately 130 years ago, effective population size was 10,000 individuals, and population size has gradually declined to 4,000 individuals. This decline does not suggest a classical founder effect, since population size remains fairly steady upon introduction to New York, but we can detect very low levels of inbreeding within some populations

(highest F_{IS} =0.082 in Washington). Importantly, the decline in N_e in the most recent time steps—the last 100 years—may be a spurious pattern resulting from known uncertainties in the final steps of this stairway plot method (Liu & Fu 2015).

Figure 3. Evidence for incipient local adaptation. A) Variation among sampling locations in environmental predictors. B) Redundancy analyses indicate that 191 SNPs (small gray points) are associated with bioclimatic predictors (vectors). BIO1: mean annual

temperature; BIO7: temperature annual range; BIO12: annual precipitation; BIO16: precipitation of wettest quarter. C) Top panel illustrates the relationship between allele frequency and strength of selection (R^2). Bottom panel assigns SNPs under selection to gene ontology categories. Colors in both panels reflect environmental predictors. D) Latent-factor mixed models show that 1315 SNPs are associated with PC1-3 of environmental variation. Black line shows 5 s.d. (black line) above the mean $\log_{10}(p\text{-value})$.

(4) Selection analyses

Starlings encounter a range of precipitation, temperature, and elevation across their range (Figure A), and about 10% of the ~15,000 SNPs sampled are associated with this environmental variation (q-value < 0.01, FDR < 0.05, TPR > 15). Redundancy analyses revealed the strongest signatures of local adaptation, showing that 191 variants are correlated with environmental differences among populations ($F = 1.022$, $P = 0.002$, Figure 3B). Populations living in warmer climates tend to cluster more closely in the left quadrant and high elevation populations cluster in the middle right quadrant. However, populations do not cluster based on geographic distance: for example, starlings from TX and WA cluster closely due to shared genetic variants, even though the two populations differ substantially in precipitation and temperature. After controlling for population structure, candidates for selection are equally distributed among elevation, temperature- and precipitation-associated predictors (Figure 3C). Mean annual temperature (BIO1) opposes selective pressure related to the range of temperatures experienced each year (BIO7), annual precipitation (BIO12), precipitation in the wettest quarter (BIO16) and elevation. Latent-factor mixed models identified 2490 candidate variants associated with the first principal component of environmental variation, which explains 41.5% of the variation and loads with temperature-related variables (Figure 3D). An additional 1315 variants were associated with precipitation-related PC2, and/or with PC3, a composite of temperature

and precipitation variables. BayeScEnv identified six SNPs (Table S2), but none of these variants were found in any other association analysis. Genes under stronger selection tend to have lower allele frequencies (Figure 3C).

Genes under putative selection function in several biological processes (Figure 3C, Table S2). Although no gene ontology categories were significantly overrepresented, signaling and response to stimuli were particularly well-represented among GO terms, showing up to 48-fold enrichment (FDR-corrected $P=0.12-0.98$, Table S2). Among signaling-related GO terms, neuron development, synaptic transmission and organization were particularly common (Table S2). Other common GO terms relate to kidney function, viral processing, metabolism, and regulation of growth factors. Among the top twenty variants under strong selection ($r^2 > 0.2$ or $\log_{10}p > 10$), we find four genes related to growth factors (EOGT, GAB3, HBEGF, STAT3), six involved in immune responses that do not directly involve growth factors (DNAJB14, FKBP4, ASB2), and three essential to muscle function (LIMCH1, HBEGF, CALD1). Putatively selected genes play a role in physiological processes that support starlings' invasion success in North America.

DISCUSSION

Our whole-genome data reveal that invasive starling populations show very low levels of genetic differentiation and moderate genetic diversity across North America. Genetic diversity is typically predicted to be low in invasive populations (Dlugosch & Parker 2008), and although RADseq tends to estimate lower genetic diversity than the true diversity, this bias is minimal for taxa known to be genetically depauperate (Cabe 1998; Cariou *et al.* 2016). Genetic variation is not explained by geographic distance, as

hierarchical AMOVAs show that variation within and among individuals explains observed differentiation better than variation among populations. These patterns are consistent with the expectation that extensive gene flow—as shown by extremely low F_{ST} among populations—maintains high connectivity across North American starling populations. Perhaps surprisingly, neither demographic reconstructions nor patterns of genetic diversity show evidence of a genetic bottleneck upon introduction, suggesting that many of the approximately 180 birds known to have been translocated went on to contribute to the invasive population.

Starlings' success depends on their ability to cope with highly variable environmental conditions across North America, and this species now thrives in habitats with dramatically different temperatures and levels of precipitation. Although there is no evidence of population structure or of isolation by distance across the North American range of this species, models indicate some subtle spatial patterns of genetic variation. Our results, combined with previous genetic studies (Cabe 1998; 1999), support the hypothesis that the Rocky Mountains may have imposed an altitudinal barrier to starlings' spread: spatially explicit models indicate a decreased migration rate on the eastern front of the mountains, and an increase in genetic diversity west of the mountain range (Figure 2). In other words, Western starling populations on the range edge are more diverse than populations nearer to the introduction site. Historical records complement this genetic evidence, as the starling expansion slowed only when reaching these mountains (Jernelov 2017). We suggest that elevation may impose a strong barrier even across this species' worldwide range: both in this study and in a parallel study of Australian populations, patterns of genetic variation can be attributed to a montane barrier (Cardilini 2016).

This genome-wide panmixia across North American starlings allows for robust tests of selection on loci that may be involved in local adaptation. Because genetic diversity and differentiation are exceptionally low, it is relatively easy to identify sites that have differentiated against the background of low genome-wide or population-specific divergence (Dlugosch *et al.* 2015). We find only six candidates out of approximately 15K SNPs using an outlier-scan method that includes environmental variation (BayeScEnv). When we control for population structure and other confounding factors (Hoban *et al.* 2016; Forester *et al.* 2018; Capblancq *et al.* 2018), we identify 191 SNPs that directly correlate with variation in temperature, precipitation, and/or elevation (Figure 3C). When controlling for geographic or other confounding variation, we find strong support for local adaptation when we explicitly model covariation among environmental and genetic distances. The relationship among allele frequency and strength of selection supports the hypothesis of incipient local adaptation, since the strongest candidates for selection tend to be rarer among North American starlings (Figure 3C).

Our study suggests that environmentally-mediated selection best explains the patterns of genetic variation observed. Although the rare variants we identify may have arisen through genetic drift, there is no evidence of genetic bottlenecks that may drive alleles towards fixation. Instead, two independent tests for genotype-environment associations show that changes in precipitation and temperature explain the low levels of genetic variation in North American starlings. Rapid evolution in genes underlying key physiological processes may have supported the starling's spread across North America. We suggest that aridity and cold temperatures that are not experienced in the starling's native range exert enough selective pressure on North American starlings to result in

incipient local adaptation. Variants in genes that coordinate physiological responses are most often correlated with precipitation in the wettest quarter (BIO16), which lends additional support to the hypothesis that arid environments which receive little rainfall year-round may generate the strongest selection on starlings. For example, effective solute transport and kidney development are critical in dry habitats, which may explain why all but one of the genes related to kidney function correlate with precipitation (BIO16). Claudin 16 (CLDN16; $R^2 = 0.23$) is one such protein that regulates ion concentrations in the kidney, while others maintain homeostasis and vasoconstriction (AVPR1B; $R^2 = 0.18$) or transport iron (STEAP3; $R^2 = 0.21$). Many invaders shift their diet upon colonization of a new habitat, and many candidates play a role in metabolism and/or digestion: for example, aridity may result in selection on proteins that process lipids (MTMR3; $R^2 = 0.18$) and fatty acids (PEX5; $R^2 = 0.17$), since organisms living in dry environments often depend on fat storage for proper hydration. Proteins that modify growth factors—key orchestrators of cellular growth and development—correlate with aridity but also temperature, and complexes that rely on ubiquitin ligases to degrade proteins are similarly strong candidates. Many of the putatively selected genes are critical to starlings' survival, and investigating a wider range of environmental conditions and sampling whole genomes may support this preliminary evidence of incipient local adaptation.

A similar study of starlings in the Australian invasion found that geographic but not environmental distance explains genetic patterns there (Cardilini 2016). Starlings in the Australian range show substantial population structuring and significant patterns of isolation-by-distance. This contrast is especially remarkable because Australian populations that are geographically close can experience dramatically different

environments, whereas environmental variation changes much more gradually across North America (Hofmeister et al., *in prep*). Because Australia is a more variable environment, starlings may have differentiated more rapidly within Australia than in North America. Global F_{ST} across all Australian populations is an order of magnitude higher than the equivalent F_{ST} index across North America, despite similar areas sampled. After controlling for the considerable structuring among Australian populations, genotype-environment associations reveal parallel signatures of selection among the two invasions. Two independent tests of incipient local adaptation in Australia and North America suggest that starlings may rapidly adapt to novel selective pressures. Both studies show that genetic variation can be explained by extremes in temperature and precipitation, and preliminary results of whole-genome resequencing of native and introduced populations confirm that variability in temperature and precipitation may shape observed genetic variation in starlings world-wide (Hofmeister et al., *in prep*).

This study explores how genetic variation changes across the landscape, but we cannot fully understand gene flow without studies of dispersal and migration of the individuals that carry genes. Juvenile dispersal promotes the high levels of gene flow observed in North America (Cabe 1999). In addition, long-distance dispersal may be common in introductions of this species: in South Africa, starlings disperse when their natal environment becomes crowded or unsuitable (Hui *et al.* 2012). In both native and introduced ranges, starlings are partial migrants: some overwintering individuals may reside in that location year-round, whereas others migrate to that wintering location and breed elsewhere (Kessel 1953). Because our samples were collected during winter, individuals within each sampling location may come from different breeding populations. If

we have only a few individuals from each of these true populations, sampling in this study may thus complicate population assignment. Although this study is the most thorough sampling of the North American range to date, additional sampling could reveal stronger population structure among true breeding populations, or divergence among resident and migrant individuals. Detailed ecological studies of dispersal and migration in North American starlings are needed to complement the genetic evidence shown here.

Our results contribute to the growing literature on rapid evolution in novel environments, even in extremely young systems. Replicate invasive populations allow us to explore the genetic consequences of colonization and establishment in novel environments. On a background of low genetic differentiation and diversity, we find evidence of incipient genotype-environment associations in North American starlings. Our results complement other recent studies that reveal associations between climate variables and particular loci in North American vertebrates (Schweizer *et al.* 2015; Bay *et al.* 2018). Finally, we suggest that rapid local adaptation can evolve even in dispersive and young populations. Studies such as this highlight the need for detailed eco-evolutionary studies of dispersal and habitat choice as potentiating forces in local adaptation.

ACKNOWLEDGEMENTS

This study would not have been possible without the sample collections provided by the U.S. Department of Agriculture's Wildlife Services personnel in Arizona, California, Colorado, Idaho, Illinois, Iowa, Kansas, Missouri, Nebraska, Nevada, New Hampshire, New Mexico, New York, North Carolina, Texas, Washington, and Wisconsin. Jennifer Walsh-Emond, Leonardo Campagna, Daniel Hooper, Jacob Berv, and Stephanie Aguillon provided valuable assistance with bioinformatic methods.

DATA ACCESSIBILITY STATEMENT

All scripts are archived on GitHub: <https://github.com/nathofme/radseq-NAm>

All data files that accompany the above scripts will be archived on Dryad upon journal submission, and DNA sequences will be archived in NCBI SRA.

REFERENCES

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403–410.
- Avalos A, Pan H, Li C *et al.* (2017) A soft selective sweep during rapid evolution of gentle behaviour in an Africanized honeybee. *Nature Communications*, 1–9.
- Bay RA, Harrigan RJ, Le Underwood V *et al.* (2018) Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science*, **359**, 83–86.
- Bitton P-P, Graham BA (2014) Change in wing morphology of the European starling during and after colonization of North America. *Journal of Zoology*, **295**, 254–260.
- Cabe PR (1998) The effects of founding bottlenecks on genetic variation in the European starling (*Sturnus vulgaris*) in North America. *Heredity*, **80**, 519–525.
- Cabe PR (1999) Dispersal and population structure in the European Starling. *The Condor*, **101**, 451–454.
- Capblancq T, Luu K, Blum MGB, Bazin E (2018) Evaluation of redundancy analysis to identify signatures of local adaptation. *Molecular Ecology Resources*, **185**, 587–11.
- Cardilini APA (2016) Phenotypic and genetic variation of non-native common starling, *Sturnus vulgaris*.
- Cariou M, Duret L, Charlat S (2016) How and how much does RAD-seq bias genetic diversity estimates? *BMC Evolutionary Biology*, 1–8.
- Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.
- Colautti RI, Lau JA (2015) Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. *Molecular Ecology*, **24**, 1999–2017.
- Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
- de Villemereuil P, Gaggiotti OE (2015) A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, **6**, 1248–1258.
- Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology*, **17**, 431–449.
- Dlugosch KM, Anderson SR, Braasch J, Cang FA, Gillette HD (2015) The devil is in the details: genetic variation in introduced populations and its contributions to invasion. *Molecular Ecology*, **24**, 2095–2111.
- Duckworth RA (2008) Adaptive Dispersal Strategies and the Dynamics of a Range Expansion. *The American Naturalist*, **172**, S4–S17.
- Earl DA, vonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.
- Edelaar P, Bolnick DI (2012) Non-random gene flow: an underappreciated force in evolution and ecology. *Trends in Ecology & Evolution*, **27**, 659–665.

- Estoup A, Ravigné V, Hufbauer R *et al.* (2016) Is There a Genetic Paradox of Biological Invasion? *Annual Review of Ecology, Evolution, and Systematics*, **47**, 51–72.
- Facon B, Hufbauer RA, Tayeh A *et al.* (2011) Inbreeding Depression Is Purged in the Invasive Insect *Harmonia axyridis*. *Current Biology*, **21**, 424–427.
- Fick SE, Hijmans RJ (2017) *WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas*. *International Journal of Climatology*.
- Forester BR, Lasky JR, Wagner HH, Urban DL (2018) Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Molecular Ecology*, **27**, 2215–2233.
- Frichot E, François O (2015) LEA: An Rpackage for landscape and ecological association studies (B O'Meara, Ed.). *Methods in Ecology and Evolution*, **6**, 925–929.
- Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, **30**, 1687–1699.
- Garant D, Forde SE, Hendry AP (2007) The multifarious effects of dispersal and gene flow on contemporary adaptation. *Functional Ecology*, **21**, 434–443.
- Guillot G, Rousset F (2013) Dismantling the Mantel tests (L Harmon, Ed.). *Methods in Ecology and Evolution*, **4**, 336–344.
- Günther T, Genetics GC, 2013 Robust identification of local adaptation from allele frequencies. *Genetics Soc America*
- .
- Hijmans RJ, van Etten J (2012) raster: Geographic analysis and modeling with raster data. R package version 2.0-12.
- Hoban S, Kelley JL, Lotterhos KE *et al.* (2016) Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, **188**, 379–397.
- Hufbauer RA, Facon B, Ravigné V *et al.* (2011) Anthropogenically induced adaptation to invade (AIAl): contemporary adaptation to human-altered habitats within the native range can promote invasions. *Evolutionary Applications*, **5**, 89–101.
- Hui C, Roura-Pascual N, Brotons L, Robinson RA, Evans KL (2012) Flexible dispersal strategies in native and non-native ranges: environmental quality and the “good-stay, bad-disperse” rule. *Ecography*, **35**, 1024–1032.
- Jacob S, Legrand D, Chaine AS *et al.* (2017) Gene flow favours local adaptation under habitat choice in ciliate microcosms. *Nature Ecology & Evolution*, 1–4.
- Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.
- Jernelov A (2017) Starlings in North America. In: *The Long-Term Fate of Invasive Species*, pp. 137–149. Jarpas.
- Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.
- Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2**, e281–14.
- Kessel B (1953) Distribution and migration of the European Starling in North America. *The Condor*, **55**, 49–67.

- Koenig W (2003) European Starlings and Their Effect on Native Cavity-Nesting Birds. *Conservation Biology*, **17**, 1134–1140.
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.
- Lenormand T (2002) Gene flow and the limits to natural selection. **17**, 183–189.
- Lescak EA, Bassham SL, Catchen J *et al.* (2015) Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences*, **112**, E7204–E7212.
- Linck EB, Battey CJ (2017) Minor allele frequency thresholds strongly affect population structure inference with genomic datasets. *bioRxiv*.
- Linz GM, Homan HJ, Gaulker SM, Penry LB, Bleier WJ (2007) *European starlings: a review of an invasive species with far-reaching impacts*.
- Liu X, Fu Y-X (2015) Exploring population size changes using SNP frequency spectra. *Nature Genetics*, **47**, 555–559.
- Lombaert E, Estoup A, Facon B *et al.* (2014) Rapid increase in dispersal during range expansion in the invasive ladybird *Harmonia axyridis*. *Journal of Evolutionary Biology*, **27**, 508–517.
- Malinsky M, Trucchi E, Lawson D, Falush D (2016) RADpainter and fineRADstructure: population inference from RADseq data. 1–6.
- Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, **28**, 659–669.
- Mi H, Dong Q, Muruganujan A *et al.* (2009) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research*, **38**, D204–D210.
- Ochocki BM, Miller TEX (2017) Rapid evolution of dispersal ability makes biological invasions faster and more variable. *Nature Communications*, **8**, 1–8.
- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species (L Orlando, Ed.). *PLoS ONE*, **7**, e37135–11.
- Petkova D, Novembre J, Stephens M (2015) Visualizing spatial population structure with estimated effective migration surfaces. *Nature Reviews Neuroscience*, **48**, 94–100.
- Pfeifer SP, Laurent S, Sousa VC *et al.* (2018) The Evolutionary History of Nebraska Deer Mice: Local Adaptation in the Face of Strong Gene Flow (N Singh, Ed.). *Molecular Biology and Evolution*, **35**, 792–806.
- Prentis PJ, Wilson JRJ, Dormontt EE, Richardson DM, Lowe AJ (2008) Adaptive evolution in invasive species. *Trends in Plant Science*, **13**, 288–294.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**, 945–959.
- Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
- Rius M, Darling JA (2014) How important is intraspecific genetic admixture to the success of colonising populations? *Trends in Ecology & Evolution*, **29**, 233–242.
- Rollins LA, Richardson MF, Shine R (2015) A genetic perspective on rapid evolution in cane toads (*Rhinella marina*). *Molecular Ecology*, **24**, 2264–2276.

- Rollins LA, Woolnough AP, Fanson BG *et al.* (2016) Selection on Mitochondrial Variants Occurs between and within Individuals in an Expanding Invasion. *Molecular Biology and Evolution*, **33**, 995–1007.
- Rollins LA, Woolnough AP, Sinclair R, MOONEY NJ, Sherwin WB (2011) Mitochondrial DNA offers unique insights into invasion history of the common starling. *Molecular Ecology*, **20**, 2307–2317.
- Rollins LA, Woolnough AP, Wilton AN, Sinclair R, Sherwin WB (2009) Invasive species can't cover their tracks: using microsatellites to assist management of starling (*Sturnus vulgaris*) populations in Western Australia. *Molecular Ecology*, **18**, 1560–1573.
- Rosenberg NA (2003) distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.
- Schlaepfer DR, Glättli M, Fischer M, van Kleunen M (2009) A multi-species experiment in their native range indicates pre-adaptation of invasive alien plant species. *New Phytologist*, **185**, 1087–1099.
- Schmid-Hempel P, Schmid-Hempel R, Brunner PC, Seeman OD, Allen GR (2007) Invasion success of the bumblebee, *Bombus terrestris*, despite a drastic genetic bottleneck. *Heredity*, **99**, 414–422.
- Schweizer RM, Robinson J, Harrigan R *et al.* (2015) Targeted capture and resequencing of 1040 genes reveal environmentally driven functional variation in grey wolves. *Molecular Ecology*, **25**, 357–379.
- Sexton JP, Hangartner SB, Hoffmann AA (2014) Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*, **68**, 1–15.
- Szűcs M, Vahsen ML, Melbourne BA *et al.* (2017) Rapid adaptive evolution in novel environments acts as an architect of population range expansion. *Proceedings of the National Academy of Sciences*, **114**, 13501–13506.
- Thomas PD, Campbell MJ, Kejariwal A *et al.* (2003) PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, **13**, 2129–2141.
- Thrasher DJ, Butcher BG, Campagna L, Webster MS, Lovette IJ (2017) Double-digest RAD sequencing outperforms microsatellite loci at assigning paternity and estimating relatedness: a proof of concept in a highly promiscuous bird. 1–34.
- Tigano A, Friesen VL (2016) Genomics of local adaptation with gene flow. *Molecular Ecology*, **25**, 2144–2164.
- Tsutsui ND, Suarez AV, Holway DA, Case TJ (2000) Reduced genetic variation and the success of an invasive species. *Proceedings of the National Academy of Sciences*, **97**, 5948–5953.
- Uller T, Leimu R (2011) Founder events predict changes in genetic diversity during human-mediated range expansions. *Global Change Biology*, **17**, 3478–3485.
- Willoughby JR, Harder AM, Tennessen JA, Scribner KT, Christie MR (2018) Rapid genetic adaptation to a novel environment despite a genome-wide reduction in genetic diversity. *Molecular Ecology*, **17**, 675.
- Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution*, **65**, 1897–1911.
- Zheng X, Levine D, Shen J *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

Table 1. Population genetic summary statistics.

Population	Private	N	P	H _{obs}	π	F _{is}
NY (10)	0	9.4	0.76	0.34	0.35	0.037
NH (11)	1	10.4	0.76	0.33	0.35	0.065
NC (11)	1	10.5	0.76	0.33	0.35	0.060
IL (7)	1	6.7	0.76	0.33	0.35	0.056
IA (9)	0	8.8	0.75	0.33	0.36	0.078
WI (8)	0	7.7	0.75	0.33	0.35	0.050
MO (9)	0	8.8	0.75	0.33	0.35	0.053
NE (11)	0	10.4	0.75	0.34	0.35	0.042
CO (8)	1	7.8	0.76	0.33	0.35	0.066
KS (9)	2	8.6	0.76	0.35	0.35	0.012
TX (9)	1	9.3	0.76	0.33	0.35	0.052
NM (7)	2	6.8	0.76	0.33	0.36	0.068
AZ (10)	1	9.5	0.76	0.32	0.35	0.069
NV (10)	0	9.5	0.75	0.33	0.36	0.068
CA (11)	0	10.4	0.75	0.33	0.35	0.061
ID (8)	0	7.7	0.76	0.33	0.35	0.047
WA (10)	2	9.5	0.76	0.32	0.35	0.082

Figure S1. This clustered coancestry matrix shows that the estimated coancestry coefficient does not vary among sampled individuals. fineRADstructure identifies fine-scale patterns in population structure as a result of shared ancestry, indicating the range of coancestry coefficients using a heat map.

Figure S2. Stairway plot indicates a gradual decline in effective population size over time. Black line indicates approximate colonization date of starlings according to historical records (1890).

742 **Table S1.** Sampling locations and details.

743

744 **Table S2.** Gene and gene ontology information.

745

746