The evolutionary history of Nebraska deer mice: local adaptation in the face of strong gene flow

Susanne P. Pfeifer[*,1,2], Stefan Laurent[*,1], Vitor C. Sousa[*,3,4], Catherine R. Linnen[*,5], Matthieu Foll[1], Laurent Excoffier[#,3], Hopi E. Hoekstra[#,6] & Jeffrey D. Jensen[#,1,2]

[*] authors contributed equally
[#] co-corresponding authors

[1] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland
[2] School of Life Sciences, Center for Evolution & Medicine, Arizona State University, USA
[3] Institute of Ecology & Evolution, University of Berne, Switzerland
[4] Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, Portugal
[5] Department of Biology, University of Kentucky, USA
[6] Department of Organismic & Evolutionary Biology and Molecular & Cellular Biology, Museum of Comparative Zoology, Howard Hughes Medical Institute, Harvard University, USA

Short title:
Local adaptation with gene flow

Corresponding authors:
Laurent Excoffier: laurent.excoffier@iee.unibe.ch
Hopi E. Hoekstra: hoekstra@oeb.harvard.edu
Jeffrey D. Jensen: jeffrey.d.jensen@asu.edu

ABSTRACT

The interplay of gene flow, genetic drift, and local selective pressure is a dynamic process that has been well studied from a theoretical perspective over the last century. Wright and Haldane laid the foundation for expectations under an island-continent model, demonstrating that an island-specific beneficial allele may be maintained locally if the selection coefficient is larger than the rate of migration of the ancestral allele from the continent. Subsequent extensions of this model have provided considerably more insight into the conditions under which such a beneficial allele may be maintained, lost, or fixed. Yet, connecting theoretical results with empirical data has proven challenging, owing to a lack of information on the relationship between genotype, phenotype, and fitness. Here, we examine the demographic and selective history of deer mice in and around the Nebraska Sand Hills, a system in which variation at the *Agouti* locus affects cryptic coloration that in turn affects the survival of mice in their local habitat. We first genotyped 250 individuals from eleven sites along a transect spanning the Sand Hills at 670,000 SNPs across the genome. Using these genomic data, we found that deer mice first colonized the Sand Hills following the last glacial period. Subsequent high rates of gene flow have served to homogenize the majority of the genome between populations on and off the Sand Hills, with the exception of the *Agouti* pigmentation locus. Furthermore, we observe strong haplotype structure around putatively beneficial mutations within the *Agouti* locus, and these mutations are strongly associated with the pigment traits that are strongly correlated with local soil coloration and thus responsible for cryptic coloration. We discuss these empirical results in light of theoretical expectations, thereby providing a complete example of the dynamics between ancestral gene flow and local adaptation in a classic mammalian system.

INTRODUCTION

Characterizing the conditions under which a population may adapt to a new environment, despite ongoing gene flow with its ancestral population, remains a question of central importance in population and ecological genetics. Hence, many studies have attempted to characterize selection-migration dynamics. Early studies focused on the distributions of phenotypes using ecological data to estimate migration rates (*e.g*., Clarke & Murray 1962; Cook & Mani 1980; Sandoval 1994); whereas later studies, focused primarily on Mendelian traits, documented patterns of genetic variation at a causal locus and contrasted that with putatively neutral loci to estimate selection and migration strengths (*e.g.,* Ross & Keller 1995; Hoekstra *et al*. 2004; Stinchcombe *et al.* 2004; and see review of Linnen & Hoekstra 2010). Only recently have such studies expanded to examine whole genome responses.

These selection-migration dynamics are well grounded in theory. Haldane (1930) first noted that migration may disrupt the adaptive process if selection is not sufficiently strong to maintain a locally beneficial allele. Bulmer (1972) went on to describe a two-alleles two-demes model in which such a loss is expected to occur if $m/s > a/(1-a)$, where m is the rate of migration, s is the selection coefficient in population 1, and a is the ratio of selection coefficients between populations. More recent related work suggests that in order for a beneficial allele to reach fixation, s must be much larger than m (*e.g*., Lenormand 2002; Yeaman & Otto 2011; and see review of Tigano & Friesen 2016).

In addition to fixation probabilities, the migration load induced by the influx of locally deleterious mutations entering the population has also been well studied. Haldane (1957) found that the number of selective deaths necessary to maintain differences between populations is proportional to the number of locally maladapted

alleles migrating into the population. Thus, the fewer loci necessary for the diverging locally adaptive phenotype, the lower the resulting migration load. More recently, Yeaman & Whitlock (2011) argued that with gene flow, the genetic architecture underlying an adaptive phenotype is expected to have fewer and larger-effect alleles compared to neutral expectations under models without migration (*i.e.*, an exponential distribution of effect sizes; Orr 1998). As recombination between these locally beneficial alleles may result in maladapted intermediate phenotypes, this model further predicts a genomic clustering of the underlying mutations (Maynard Smith 1977; Lenormand & Otto 2000). The relative advantage of this linkage will increase as the population approaches the migration threshold, at which local adaptation becomes impossible (Kirkpatrick & Barton 2006). Taken together, these studies make a number of testable predictions relating migration rate with the expected number of loci underlying the locally beneficial phenotype, the average effect size of the alleles, the clustering of beneficial mutations, and the conditions under which a locally beneficial allele may be lost, maintained, or fixed in a population.

These theoretical expectations, however, have proven difficult to evaluate in natural populations owing to a dearth of systems with concrete links between genotype, phenotype, and fitness. In one well-studied system, deer mice (*Peromyscus maniculatus*) inhabiting the light-colored Sand Hills of Nebraska have evolved lighter coloration than conspecifics on the surrounding dark soil (Dice 1941, 1947). Using a combination of laboratory crosses and hitchhiking-mapping, the *Agouti* gene, which encodes a signaling protein known to play a key role in mammalian pigment-type switching and color patterning (Jackson 1994; Barsh 1996; Mills & Patterson 2009; and see review of Manceau *et al.* 2010), has been implicated in adaptive color variation (Linnen *et al*. 2009). Moreover, using association mapping, specific

candidate mutations have been linked to variation in specific pigment traits, which in turn contribute to differential survival from visually hunting avian predators (Linnen *et al*. 2013).

While previous research in this system has established links between genotype, phenotype, and fitness, this work has focused on a single ecotonal population located near the northern edge of the Sand Hills. Thus, the dynamic interplay of migration, selection, and genetic drift, as well as the extent of genotypic and phenotypic structuring among populations, remains unknown. To address these questions, we have sampled hundreds of individuals across a transect spanning the Sand Hills and neighboring populations to the north and south. Combining extensive per-locale soil color measurements, per-individual phenotyping, and 670,000 SNPs genome-wide, this dataset provides a unique opportunity to evaluate theoretical predictions of local adaptation with gene flow on a genomic scale.

RESULTS & DISCUSSION

Environmental and phenotypic variation in and around the Nebraska Sand Hills

We collected *Peromyscus maniculatus luteus* individuals ($N = 266$) as well as soil samples ($N = 271$) from 11 locations spanning a 330 km transect across the Sand Hills of Nebraska (Figure 1; Supplementary Table 1). As expected, we found that soil color differed significantly between "on" Sand Hills and "off" Sand Hills locations ($F_1 = 307.4$, $P < 1 \times 10^{-15}$; Figure 2). Similarly, five of the six mouse color traits also differed significantly between the two habitats (Figure 2). In contrast, mice trapped "on" versus "off" of the Sand Hills did not differ in other non-pigmentation traits including total body length, tail length, hind foot length, or ear length. Together, these

results are consistent with strong divergent selection between habitat types on multiple color traits, but not other morphological traits.

We also examined habitat heterogeneity and phenotypic differences among sampling sites within habitat types. Both soil brightness and mouse color differed significantly among sites within habitats (soil: $F_9 = 8.0$, $P = 1.9 \times 10^{-10}$; dorsal brightness: $F_9 = 10.5$, $P = 8.1 \times 10^{-14}$; dorsal hue: $F_9 = 3.4$, $P = 6.7 \times 10^{-4}$; ventral brightness: $F_9 = 10.3$, $P = 1.4 \times 10^{-13}$; dorsal-ventral boundary: $F_9 = 6.6$, $P = 1.5 \times 10^{-8}$; tail stripe: $F_9 = 6.9$, $P = 7.5 \times 10^{-9}$). Additionally, site-specific means for three of the color traits were significantly correlated with local soil brightness, including: dorsal brightness ($t = 4.4$, $P = 0.0017$), dorsal-ventral boundary ($t = 3.9$, $P = 0.0034$), and tail stripe ($t = 3.1$, $P = 0.013$). In contrast, site-specific means for dorsal hue and ventral brightness did not correlate with local soil color (dorsal hue: $t = 2.385$, $P = 0.097$; ventral brightness: $t = 0.32$; $P = 0.77$). These results suggest that the agents of selection shaping correlations between local soil color and mouse color vary among the traits.

The genetic architecture of light color in Sand Hills mice

Genetic architecture parameter estimates derived from our association mapping analyses suggest that variants in the *Agouti* locus explain a considerable amount of the observed color variation in on versus off Sand Hills populations (Table 1), although other pigmentation loci likely contribute as well. For example, *Agouti* SNPs explain 69% of the total phenotypic variance in dorsal brightness and tail stripe. This analysis also provides estimates of the number of causal SNPs as well as the proportion of genetic variance that is attributable to major-effect SNPs. These estimates suggest that potentially dozens of *Agouti* variants may be contributing to

variation in color traits, but the credible intervals for SNP number are wide (Table 1). One explanation for such wide intervals is that it is difficult to disentangle the effects of closely linked SNPs, thus this number is likely an over-estimate. Nevertheless, because the lower bound of the credible intervals for all but one color trait exceeds one, our data indicate that there are multiple *Agouti* mutations with non-negligible effects on color phenotypes. Estimates of PGE (percent of genetic variance attributable to major effect mutations) indicate that, whatever their number, these major effect mutations explain a high percentage of total genetic variance (*e.g.,* 83% for dorsal brightness and 87% for dorsal-ventral boundary).

Table 1. Hyper-parameters describing the genetic architecture of five color traits. All parameters were estimated using a BSLMM model in GEMMA, *Agouti* SNPs, and a relatedness matrix estimated from genome-wide SNPs. Parameter means and 95% credible intervals (lower bound, upper bound) were averaged across ten independent runs per trait. Interpretation of hyper-parameters is as follows: *PVE,* the total proportion of phenotypic variance that is explained by genotype; *PGE,* the proportion of genetic variance explained by sparse (*i.e.*, major) effects; *h,* an approximation of the expected value of *PVE*; *rho,* an approximation of the expected value of *PGE*; *n_gamma,* the expected number of sparse (major) effect SNPs.

Trait	h	PVE	rho	PGE	n_gamma
Dorsal brightness	0.50 (0.29, 0.7)	0.69 (0.57, 0.81)	0.62 (0.31, 0.92)	0.83 (0.64, 0.97)	117.2 (7.6, 286)
Dorsal hue	0.28 (0.07, 0.57)	0.26 (0.08, 0.45)	0.32 (0.02, 0.87)	0.30 (0, 0.86)	54.4 (0, 262.7)
Ventral brightness	0.33 (0.16, 0.51)	0.39 (0.23, 0.56)	0.43 (0.13, 0.82)	0.54 (0.22, 0.88)	80.6 (3, 269.9)
D-V boundary	0.44 (0.25, 0.64)	0.60 (0.46, 0.73)	0.75 (0.45, 0.98)	0.87 (0.7, 0.99)	70.5 (9.1, 253.7)
Tail stripe	0.51 (0.32, 0.69)	0.69 (0.56, 0.8)	0.64 (0.38, 0.89)	0.79 (0.61, 0.95)	123.9 (16.6, 283.7

Although our genetic architecture parameter estimates indicate that large-effect *Agouti* variants contribute to variation in each of the five color traits, the number and location of these SNPs vary considerably (Figure 3). To interpret these

results, some background on the structure and function of *Agouti* is needed. Work in *Mus musculus* has identified two different *Agouti* isoforms that are under the control of different promoters. The ventral isoform contains noncoding exons 1A/1A', is expressed in the ventral dermis during embryogenesis and is required for determining the boundary between the dark dorsum and light ventrum (Bultman *et al.* 1994; Vrieling *et al.* 1994). The hair-cycle isoform, containing noncoding exons 1B/1C, is expressed in both the dorsal and ventral dermis during hair growth, and is required for forming light, pheomelanin bands on individual hairs (Bultman *et al.* 1994; Vrieling *et al.*1994). In *Peromyscus*, these same isoforms are present as well as additional, novel isoforms (Mallarino *et al*. 2017). Isoform-specific changes in *Agouti* expression are associated with changes in the dorsal-ventral boundary (ventral isoform; Manceau *et al.* 2011) and the width of light bands on individual hairs (hair-cycle isoform; Linnen *et al.* 2009). Both isoforms share the same three coding exons (exons 2, 3, and 4); thus, protein-coding changes could simultaneously alter pigmentation patterning and hair banding (Linnen *et al.* 2013).

Across the *Agouti* locus, we detected 160 SNPs that were strongly associated (posterior inclusion probability [PIP] in the polygenic BSLMM > 0.1) with at least one color trait. By trait, the number of SNPs exceeding our PIP threshold was: 53 (dorsal brightness), 2 (dorsal hue), 16 (ventral brightness), 34 (dorsal-ventral boundary), and 81 (tail stripe). Additionally, patterns of genotype-phenotype association were distinct for each trait but consistent with *Agouti* isoform function (Figure 3). For example, we observed the strongest associations for dorsal brightness around the three coding exons (exons 2, 3, and 4, located between positions 25.06Mb and 25.11Mb on the scaffold containing the *Agouti* locus) and upstream of the transcription start site for the *Agouti* hair-cycle isoform (Figure 3). By contrast, for

dorsal-ventral boundary, the SNP with the highest PIP was located very near the transcription start site for the *Agouti* ventral isoform (Figure 3). Notably, a previously identified serine deletion in exon 2 (Linnen et al. 2009; 2013) exceeds our threshold for three of the color traits (dorsal brightness, dorsal-ventral boundary, and tail stripe) and was elevated above background association levels for the remaining two traits (dorsal hue and ventral brightness); and multiple candidates from those analyses are also recovered here (Supplementary Table 2). Overall, our association mapping results reveal that there are many sites in *Agouti* that are associated with pigment variation.

The demographic history of the Sand Hills population

Based on their genetic diversity, individuals from different sampling locations grouped according to their location and habitat, with a clear separation between individuals from on and off of the Sand Hills (Figure 4a). The observed pattern of isolation by distance was further supported by a significant correlation between pairwise genetic differentiation ($F_{ST}/(1-F_{ST})$) and pairwise geographic distance among sampling sites ($P = 9.9 \times 10^{-4}$; Figure 4b). The pairwise F_{ST} values between sampling locations were very low, ranging from 0.008 to 0.065, indicating limited genetic differentiation (Figure 4c). Consistent with both patterns of isolation by distance and reduced differentiation among populations, individual ancestry proportions (as inferred by sNMF) were best explained by three clusters. Additionally, the ancestry proportions obtained with three distinct clusters (Figure 4d) returned low cross-entropy values (Supplementary Figure 1), again suggesting three distinct groups: (i) north of the Sand Hills, (ii) on the Sand Hills, and (iii) south of the Sand Hills. The population tree inferred with TreeMix also indicated low levels of genetic drift among

populations, but a pattern consistent with increased differentiation moving along the transect (Figure 4e).

Given the observed population structure, we next investigated models of colonization history based on three populations: those inhabiting the Sand Hills and those inhabiting the neighboring regions to the north and to the south. This resulted in two different three-population demographic models, corresponding to different population tree topologies, which explicitly account for bottlenecks associated with potential founder events and gene flow among populations (Supplementary Figure 2). Both models appeared equally likely (Supplementary Table 3), and parameter estimates were similar and consistent across models, pointing to a recent divergence time among populations, evidence of a bottleneck associated with the colonization of the Sand Hills, and high rates of migration among all populations (Supplementary Table 4).

To better distinguish between the models, we compared the likelihoods of the two models for bootstrap data sets containing a single SNP per 1.5kb block, counting the number of bootstrap replicates with a relative likelihood (based on AIC values) larger than 0.95 for each model. Using this approach, we identified a model with a topology in which the population on the Sand Hills shares a more recent common ancestor with the population off the Sand Hills to the south (supported in 81/100 bootstrap replicates, Supplementary Figure 3). This topology and the parameter estimates (Supplementary Table 4) are consistent with a recent colonization of the Sand Hills from the south, namely: (i) a recent split within the last ~4,000 years (95% CI 3,400 to 7,900); (ii) a stronger bottleneck associated with the colonization of the Sand Hills compared to the older split between northern and southern populations; and (iii) higher levels of gene flow from south to north, including higher migration

rates from the southern population in to the Sand Hills (2*Nm* 95% CI: 12.5-24.0). Furthermore, this model fits well the marginal site frequency spectrum (SFS) for each sample as well as the 2D marginal SFS for each pair of populations (Supplementary Figure 4). Thus, this neutral demographic model nicely explains observed patterns of variation outside of the *Agouti* region.

The selective history of the Sand Hills population

Despite the high levels of gene flow inferred in our model, which results in low levels of differentiation among populations genome-wide (Supplementary Figure 5a), we observed high levels of differentiation among populations at the *Agouti* locus – variation which is further correlated with different phenotypic trait values (Figure 5; Supplementary Figures 5,6). We observed the highest level of differentiation between mice sampled on either side of the northern limit of the Sand Hills, with a 100kb region within *Agouti* being the largest observed F_{ST} value in our dataset (Supplementary Figure 6). This pattern is consistent with the expectations of positive selection under a local adaptation regime. Within this 100kb region, a sub-region of 30kb located in intron 1 (*i.e.,* between exon 1A and exon 1) displayed maximal differentiation (Supplementary Figure 6), making it difficult to precisely identify the target of selection. In contrast to the wide-range differentiation observed at *Agouti* on the northern side of the cline, a single narrow F_{ST}-peak of roughly 5kb located in intron 1 was observed on the southern limit of the Sand Hills (Figure 5; Supplementary Figure 6a). Genome-wide comparison confirmed the overall genetic similarity between light-colored Sand Hills mice and the dark-colored population to the south, with the exception of a small number of variants putatively driving adaptive phenotypic divergence (Figure 5; Supplementary Figure 6a).

To test whether this pattern of differentiation could be the result of non-selective forces (see Crisci *et al.* 2012; Jensen *et al.* 2016), we calculated the HapFLK statistic, providing a single measure of genetic differentiation for the three geographic localities while controlling for neutral differentiation. Before calculating genetic differentiation at the target locus (*i.e., Agouti*), the HapFLK method computes a neutral distance matrix from the background data (*i.e.,* the anonymous regions randomly distributed across the genome) that summarizes the genetic similarity between populations with regard to the extent of genetic drift since divergence from their common ancestral population (Supplementary Figure 7). Consistent with the inferred demographic history of the populations, genome-wide background levels of variation suggest that individuals captured south of the Sand Hills are more similar to those inhabiting the Sand Hills, compared with populations to the north. The HapFLK profile confirmed the significant genetic differentiation at the *Agouti* locus compared to neutral expectations (Figure 6). Interestingly, southern populations share a significant amount of the haplotypic structure observed in populations on the Sand Hills, with the exception of the candidate region itself. This further indicates that high rates of on-going gene flow occur between on and off Sand Hills populations, but that selection maintains the ancestral *Agouti* alleles in the populations off the Sand Hills, and the derived alleles conferring crypsis on the Sand Hills.

Owing to low levels of linkage disequilibrium characterizing the *Agouti* region, adaptive variants could be mapped on a fine genetic scale. Specifically, there are three regions of increased differentiation (Figure 6): (i) the narrow 3kb peak in the HapFLK profile (located in intron 1; Supplementary Table 5) co-localizes with a region of high linkage disequilibrium (Supplementary Figure 8), suggesting a recent selective event; (ii) the second highest peak of differentiation, located between exon

1A and the duplicated reversed copy, exon1A', is the only significant region detected by the CLR test (Supplementary Figure 9) and contains strong haplotype structure (Figure 6), again suggesting the recent action of positive selection; and (iii) the third highest peak of differentiation surrounds the putatively beneficial serine deletion in exon 2 previously described by Linnen *et al.* (2009), showing strong differentiation between on and off Sand Hills populations in our dataset. Although several lines of evidence support the role of the serine deletion in adaptation to the Sand Hills, the linkage disequilibrium around this variant is low, suggesting that the age of the corresponding selective event is likely the most ancient amongst the three candidate regions, with subsequent mutation and recombination events reducing the selective sweep pattern.

Predicted migration thresholds and the conditions of allele maintenance

Although a number of strong assumptions must be made, it is possible to estimate the migration threshold below which a locally beneficial mutation may be maintained in a population. Given our high rates of inferred gene flow, it is indeed important to consider whether observations are consistent with these theoretical expectations necessary for the maintenance of alleles. Following Yeaman & Whitlock (2011) and Yeaman & Otto (2011), this migration threshold is defined in terms of both rates of gene flow as well as the fitness effects of each respective locally beneficial allele in the matching and alternate environment. Given the parameter values inferred here, along with estimated population sizes, the threshold above which the most strongly beneficial mutations may be maintained in the population is estimated at $m = 0.8$ – a large value indeed, given the inferred strength of selection. For the more weakly beneficial mutations identified, this value is estimated at $m =$

0.07, again well above empirically estimated rates. Thus, our empirical observations are fully consistent with expectation in that gene flow is strong enough to prevent locally beneficial alleles from fixing on the Sand Hills, but not so strong so as to swamp out the derived allele despite the high input of ancestral variation.

CONCLUSIONS

The cryptically colored mice of the Nebraska Sand Hills represent one of the few mammalian systems in which aspects of genotype, phenotype, and fitness have been measured and connected. Yet, the population genetics of the Sand Hills region has remained poorly understood. By sequencing hundreds of individuals across a cline spanning over 300 km, fundamental aspects of the evolutionary history of this population could be addressed for the first time. Utilizing genome-wide putatively neutral regions, the inferred demographic history suggests a colonization of the Sand Hills from the south following the last glacial period. Further, high rates of gene flow are inferred between light and dark populations - resulting in low levels of genetic differentiation genome-wide as well as low levels of phenotypic differentiation of traits (unrelated to coloration) across the cline. However, the *Agouti* region differs markedly in this regard, with high levels of differentiation observed between on and off Sand Hills populations, strong haplotype structure, and high levels of linkage disequilibrium spanning putatively beneficial mutations among cryptically colored individuals. In addition, we found these putatively beneficial mutations to be strongly associated with the phenotypic traits underlying crypsis, and these phenotypic traits were found to be in strong association with variance in soil color across the cline.

Together, these results suggest a model in which selection is acting to maintain the alleles underlying the locally adaptive phenotype on light/dark soil, despite

substantial gene flow, which not only prevents the populations from strongly differentiating from one another, but also prevents the cryptic genotypes from reaching local fixation. Furthermore, consistent with theoretical expectations, the mutations underlying the cryptic phenotype are found to be few in number, of large effect size, and in close genomic proximity. As described by Yeaman & Whitlock (2011), at the outset of local adaptation under models with high migration, the establishment of a large effect beneficial allele may indeed facilitate the accumulation of other locally beneficial alleles in the same genomic region owing to the local reduction in effective migration rate (and see Aeschbacher & Burger 2014). Our results are in keeping with this model, with an exonic deletion explaining a large proportion of variance in multiple traits underlying the cryptic phenotype, and with population-genetic patterns suggesting comparatively ancient selection acting on this variant. In addition, multiple genomically clustered and smaller effect alleles modulating individual traits are also identified, which are characterized by strong patterns of selective sweeps, indicative of more recent selection. This system thus provides an in-depth picture of the dynamic interplay of these processes in a natural mammalian population, characterized by remarkably strong local selective pressures as well as continuous high rates of gene flow with the ancestral founding populations.

MATERIAL AND METHODS

Population Sampling

We collected *Peromyscus maniculatus luteus* individuals ($N = 266$) and corresponding soil samples ($N = 271$) from eleven sites spanning a 330 km transect starting ~120 km north of the Sand Hills and ending ~120 km south of the Sand Hills (Figure 1; Supplementary Table 1). In total, five sampling locations were on the Sand

Hills and six were located off of the Sand Hills (three in the north and three in the south). We collected mice using Sherman live traps and prepared flat skins according to standard museum protocols; these specimens are accessioned in the Harvard University Museum of Comparative Zoology's Mammal Department. From each sample, we preserved liver tissue in 100% EtOH for subsequent DNA extraction. Collections were made under the South Dakota Department of Game, Fish, and Parks scientific collector's permit 54 (2008) and the Nebraska Game and Parks Commission Scientific and Educational Permit (2008) 579, sub-permit 697-700. This work was approved by Harvard University's Institutional Animal Care and Use Committee.

Mouse color measurements: For each mouse, we measured standard morphological traits, including: total length, tail length, hind foot length, and ear length. We also characterized color and color pattern using methods modified from Linnen *et al.* (2013). In brief, to quantify color, we used a USB2000 spectrophotometer (Ocean Optics) to take reflectance measurements at three sites across the body (three replicated measurements in the dorsal stripe, flank, and ventrum). We processed these raw reflectance data using the CLR: Colour Analysis Programs v.1.05 (Montgomerie 2008). Specifically, we trimmed and binned the data to 300-700nm and 1nm bins and computed five color summary statistics: B2 (mean brightness), B3 (intensity), S3 (chroma), S6 (contrast amplitude), and H3 (hue). We then averaged the three measurements for each body region, producing a total of 15 color values (*i.e.,* five summary statistics from each of the three body regions). To ensure values were normally distributed, we performed a normal-quantile transformation on each of the 15 color variables. Using these transformed values, we performed a principal component analysis (PCA) in STATA v.13.0 (StataCorp, College Station, TX). Based

on eigenvalues and examination of the scree plot, we selected the first four principal components, which together accounted for 84% of the variation in color. To increase interpretability of the loadings, we performed a VARIMAX rotation on the first four PCs (Tabachnick & Fidell 2001; Montgomerie 2006). After rotation, PC1 corresponded to the brightness/contrast (B2, B3, S6) of the dorsum; PC2 to the chroma/hue (S3, H3) of the dorsum; PC3 to the brightness/contrast (B2, B3, S6) of the ventrum; and PC4 to the chroma/hue (S3, H3) of the ventrum. Therefore, we refer to PC1 as "dorsal brightness", PC2 as "dorsal hue", PC3 as "ventral brightness", and PC4 as "ventral hue" throughout.

To quantify color pattern, we took digital images of each mouse flat skin with a Canon EOS Rebel XTI with a 50mm Canon lens (Canon U.S.A., Lake Success, NY). We used the quick selection tool in Adobe Photoshop CS5 (Adobe Systems, San Jose, CA) to select light and dark areas on each mouse. Specifically, we outlined the dorsum (brown portion of the dorsum with legs and tails excluded), body (outlined entire mouse with brown and white areas, legs and tails excluded), tail stripe (outlined dark stripe on tail only), and tail (outlined entire tail). We calculated "dorsal-ventral boundary" as (body – dorsum)/body and "tail stripe" as (tail – tail stripe)/tail. Thus, each measure represents the proportion of a particular region (tail or body) that appeared unpigmented; higher values therefore represent lighter phenotypes.

To determine whether color phenotypes (dorsal brightness, dorsal hue, ventral brightness, ventral hue, dorsal-ventral boundary, and tail stripe) differ between on Sand Hills and off Sand Hills populations, we performed a nested analysis of variance (ANOVA) on each trait, with sampling site nested within habitat. For comparison, we also performed nested ANOVAs on each of the four non-color traits (total length, tail length, hind foot length, and ear length). All analyses were performed on normal-

quantile-transformed data. Unless otherwise noted, we performed all transformations and statistical tests in R v3.3.2.

Soil color measurements: To characterize soil color, we collected soil samples in the immediate vicinity of each successful *Peromyscus* capture. To measure brightness, we poured each soil sample into a small petri dish and recorded ten measurements using the USB2000 spectrophotometer. As above, we used CLR to trim and bin the data to 300-700nm and 1nm bins as well as to compute mean brightness (B2). We then averaged these ten values to produce a single mean brightness value for each soil sample, and transformed the soil-brightness values using a normal-quantile transformation prior to analysis. To evaluate whether soil color differs consistently between on and off of the Sand Hills sites, we performed an ANOVA, with sampling site nested within habitat. Finally, to test for a correlation between color traits (see above) and local soil color, we regressed the mean brightness value for each color trait on the mean soil brightness for each sampling site – ultimately for comparison with downstream population genetic inference (see Joost *et al*. 2013).

Library Preparation and Sequencing

Library Preparation: To prepare sequencing libraries, we used DNeasy kits (Qiagen, Germantown, MD) to extract DNA from liver samples. We then prepared libraries following the SureSelect Target Enrichment Protocol v.1.0, with some modifications. In brief, 10-20 μg of each DNA sample was sheared to an average size of 200bp using a Covaris ultrasonicator (Covaris Inc., Woburn, MA). Sheared DNA samples were purified with Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN) and quantified using a Quant-iT dsDNA Assay Kit (ThermoFisher Scientific, Waltham,

MA). We then performed end repair and adenylation, using 50 μl ligations with Quick Ligase (New England Biolabs, Ipswich, MA) and paired-end adapter oligonucleotides manufactured by Integrated DNA Technologies (Coralville, IA). Each sample was assigned one of 48 five-base pair barcodes. We pooled samples into equimolar sets of 9-12 and performed size selection of a 280-bp band (+/- 50bp) on a Pippin Prep with a 2% agarose gel cassette. We performed 12 cycles of PCR with Phusion High-Fidelity DNA Polymerase (NEB), according to manufacturer guidelines. To permit additional multiplexing beyond that permitted by the 48 barcodes, this PCR step also added one of 12 six-base pair indices to each pool of twelve barcoded samples. Following amplification and AMPure purification, we assessed the quality and quantity of each pool (23 total) with an Agilent 2200 TapeStation (Agilent Technologies, Santa Clara, CA) and a Qubit-BR dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA).

Enrichment and Sequencing: To enrich sample libraries for both the *Agouti* locus as well as ~1000 randomly distributed genomic regions, we used a MYcroarray MYbaits capture array (MYcroarray, Ann Arbor, MI). This probe set includes the 185kb region containing all known *Agouti* coding exons and regulatory elements (based on a *P. maniculatus rufinus Agouti* BAC clone from Kingsley *et al.* 2009) and >1000 non-repetitive regions averaging 1.5kb in length at random locations across the *P. maniculatus* genome.

After generating 23 indexed pools of barcoded libraries from 249 individual Sand Hills mice and one lab-derived non-agouti control, we enriched for regions of interest following the standard MYbaits v.2 protocol for hybridization, capture, and recovery. We then performed 14 cycles of PCR with Phusion High-Fidelity Polymerase and a final AMPure purification. Final quantity and quality was then

assessed using a Qubit-HS dsDNA Assay Kit and Agilent 2200 TapeStation. After enriching our libraries for regions of interest, we combined them into four pools and sequenced across eight lanes of 125bp paired-end reads using an Illumina HiSeq2500 platform (Illumina, San Diego, CA). Of the read pairs, 94% could be confidently assigned to individual barcodes (*i.e.,* we excluded read data where the ID tags were not identical between the two reads of a pair (4%) as well as reads where ID tags had low base qualities (2%)). Read pair counts per individual ranged from 367,759 to 42,096,507 (median 4,861,819).

Reference Assembly

We used the *Peromyscus maniculatus bairdii* Pman_1.0 reference assembly publicly available from NCBI (RefSeq assembly accession GCF_000500345.1), which consists of 30,921 scaffolds containing 2,630,541,020bp ranging from 201bp to 18,898,765bp in length (median scaffold length: 2,048bp; mean scaffold length: 85,073bp) to identify variation in the background regions (*i.e.*, outside of *Agouti*). In this assembly, the *Agouti* locus is split over two scaffolds (*i.e.*, exon 1A/A' is located on scaffold NW_006502894.1 and exon 2, 3, and 4 are located on scaffold NW_006501396.1), thus another in-house *Peromyscus* reference assembly was used to identify variation in *Agouti*. This reference assembly consists of 9,080 scaffolds containing 2,512,380,343bp ranging from 1,000bp to 60,475,073bp in length (median scaffold length: 3,275bp; mean scaffold length: 276,694bp). In the two reference assemblies, we annotated and masked seven different classes of repeats (*i.e.,* SINE, LINE, LTR elements, DNA elements, satellites, simple repeats, and low complexity regions) using RepeatMasker v.Open-4.0.5 (Smit *et al.* 2013-2015).

Sequence Alignment

We removed contamination from the raw read pairs and trimmed low quality read ends using cutadapt v.1.8 (Martin 2011) and TrimGalore! v.0.3.7 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore). We aligned the preprocessed, paired-end reads to the reference assembly using Stampy v.1.0.22 (Lunter & Goodson 2011). PCR duplicates as well as reads that were not properly paired were then removed using SAMtools v.0.1.19 (Li *et al.* 2009). After cleaning, read pair counts per individual ranged from 220,960 to 20,178,669 (median: 3,062,998). We then conducted a multiple sequence alignment using the Genome Analysis Toolkit (GATK) IndelRealigner v.3.3 (McKenna *et al.* 2010; DePristo *et al.* 2011; Van der Auwera *et al.* 2013) to improve variant calls in low-complexity genomic regions, adjusting Base Alignment Qualities (BAQ) at the same time in order to down weight base qualities in regions with high ambiguity (Li 2011). Next, we merged sample-specific reads across different lanes, thereby removing optical duplicates using SAMtools v.0.1.19. Following GATK's Best Practice, we performed a second multiple sequence alignment to produce consistent mappings across all lanes of a sample. The resulting dataset contained aligned read data for 249 individuals from eleven different sampling locations.

Variant Calling and Filtering

We performed an initial variant call using GATK's HaplotypeCaller v.3.3 (McKenna *et al.* 2010; DePristo *et al.* 2011; Van der Auwera *et al.* 2013), and then we genotyped all samples jointly using GATK's GenotypeGVCFs v.3.3 tool. Post-genotyping, we filtered initial variants using GATK's VariantFiltration v.3.3, removing SNPs using the following set of criteria (with acronyms as defined by the

GATK package): (i) the variant exhibited a low read mapping quality (MQ<40); (ii) the variant confidence was low (QD<2.0); (iii) the mapping qualities of the reads supporting the reference allele and those supporting the alternate allele were qualitatively different (MQRankSum<12.5); (iv) there was evidence of a bias in the position of alleles within the reads that support them between the reference and alternate alleles (ReadPosRankSum<-8.0); or (v) there was evidence of a strand bias as estimated by Fisher's exact test (FS>60.0) or the Symmetric Odds Ratio test (SOR>4.0). We removed indels using the following set of criteria: (i) the variant confidence was low (QD<2.0); (ii) there was evidence of a strand bias as estimated by Fisher's exact test (FS>200.0); or (iii) there was evidence of bias in the position of alleles within the reads that support them between the reference and alternate alleles (ReadPosRankSum<-20.0). For an in-depth discussion of these issues, see the recent review of Pfeifer (2017).

To minimize genotyping errors, we excluded all variants with a mean genotype quality of less than 20 (corresponding to P[error] = 0.01). We limited SNPs to biallelic sites using VCFtools v.0.1.12b (Danecek *et al.* 2011), with the exception of one previously identified putatively beneficial triallelic variant. We excluded variants within repetitive regions of the reference assembly from further analyses as potential mis-alignment of reads to these regions might lead to an increased frequency of heterozygous genotype calls. Additionally, we filtered variants on the basis of the Hardy-Weinberg equilibrium (HWE) by computing a *p*-value for HWE for each variant using VCFtools v.0.1.12b and excluding variants with an excess of heterozygous ($p < 0.01$), unless otherwise noted. Finally, we removed sites for which all individuals were fixed for the non-reference allele as well as sites with more than 25% missing genotype information.

The resulting call set contained 8,265 variants on scaffold 16 (containing *Agouti*) and 651,408 variants within the random background regions. We polarized variants within *Agouti* using the *P. maniculatus rufinus Agouti* sequence (Kingsley *et al.* 2009) as the outgroup. Genotypes were phased using BEAGLE v.4 (Browning & Browning 2007).

Accessible Genome

To minimize the number of false positives in our dataset, we subjected variants to several stringent filter criteria. The application of these filters led to an exclusion of a substantial fraction of genomic regions, and thus we generated mask files (using the GATK pipeline described above, but excluding variant-specific filter criteria *i.e.*, QD, FS, SOR, MQRankSum, and ReadPosRankSum) to identify all nucleotides accessible to variant discovery. After filtering, only a small fraction of the genome (*i.e.*, 63,083bp within *Agouti* and 6,224,355bp of the random background regions) remained accessible. These mask files enabled us to obtain an exact number of the monomorphic sites in the reference assembly, which we then used to both estimate the demographic history of focal populations as well as a control to avoid biases when calculating summary statistics.

Association mapping

To identify SNPs within *Agouti* that contribute to variation in mouse coat color, we used the Bayesian sparse linear mixed model (BSLMM) implemented in the software package GEMMA v 0.94.1 (Zhou *et al.* 2013). In contrast to single-SNP association mapping approaches (*e.g.,* Purcell *et al.* 2007), the BSLMM is a polygenic model that simultaneously assesses the contribution of multiple SNPs to phenotypic

variation. Additionally, compared to other polygenic models (*e.g.,* linear mixed models and Bayesian variable selection regression), the BSLMM performs well for a wide range of trait genetic architectures, from a highly polygenic architecture with normally distributed effect sizes to a "sparse" model in which there are a small number of large-effect mutations (Zhou *et al.* 2013). Indeed, one benefit of the BSLMM (and other polygenic models) is that hyper-parameters describing trait genetic architecture (*e.g.,* number of SNPs, relative contribution of large- and small-effect variants) can be estimated from the data. Importantly, the BSLMM also accounts for population structure via inclusion of a kinship matrix as a covariate in the mixed model.

For each of the five color traits that differed significantly between on Sand Hills and off Sand Hills habitats (*i.e.*, dorsal brightness, dorsal hue, ventral brightness, dorsal-ventral boundary, and tail stripe, see "Results and Discussion"), we performed ten independent GEMMA runs, each consisting of 25 million generations, with the first five million generations discarded as burn-in. Because we were specifically interested in the contribution of *Agouti* variants to color variation, we restricted our GEMMA analysis to 2,148 *Agouti* SNPs that had no missing data and a minor allele frequency (MAF) > 0.05. However, to construct a relatedness matrix, we used the genome-wide SNP dataset. We aimed to maximize the number of individuals kept in the analyses, and hence we used less stringent filtering criteria than for the demographic analyses. Prior to construction of this matrix, we removed all individuals with a mean depth (DP) of coverage lower than 2X. Following the removal of low-coverage individuals, we treated genotypes with less than 2X coverage or twice the individual mean DP as missing data and removed SNPs with more than 10% missing data across individuals or with a MAF<0.01. To remove tightly linked SNPs, we

sampled one SNP per 1kb block, choosing whichever SNP had the lowest amount of missing data, resulting in a dataset with 12,920 SNPs (Supplementary Table 6). We then computed the relatedness matrix using the R/Bioconductor (release 3.4) package SNPrelate v1.8.0 (Zheng *et al.* 2012). For all traits, we used normal quantile transformed values to ensure normality, and option "-bslmm 1" to fit a standard linear BSLMM to our data.

After runs were complete, we assessed convergence on the desired posterior distribution via examination of trace plots for each parameter and comparison of results across independent runs. We then summarized posterior inclusion probabilities (PIP) for each SNP for each trait by averaging across the ten independent runs. Following Chaves *et al.* (2016), we used a strict cut-off of PIP>0.1 to identify candidate SNPs for each color trait (cf. Gompert *et al.* 2013; Comeault *et al.* 2014). To summarize hyper-parameter estimates for each trait, we calculated the mean and upper and lower bounds of the 95% confidence interval for the parameter posterior distributions obtained in each run and then averaged these values across the ten independent runs.

Population Structure

We investigated the structure of populations along the Nebraska Sand Hills transect with several complementary approaches. First, we used methods to cluster individuals based on their genetic similarity, including PCA and inference of individual ancestry proportions. Second, on the basis of SNP allele frequencies, we computed pairwise F_{ST} between sampled sites to infer their relationships. In addition, we inferred the population tree best fitting the covariance of allele frequencies across sites using TreeMix v1.13 (Pickrell & Pritchard 2012). Finally, we tested for isolation

by distance (IBD) by comparing the matrix of geographic distances between sites to that of pairwise genetic differentiation ($F_{ST}/(1-F_{ST})$; Rousset 1997) using Mantel tests (Sokal 1979; Smouse *et al.* 1986).

For demographic analyses, we based all analyses on the anonymous background regions randomly distributed across the genome (*i.e.,* excluding *Agouti*) and applied additional filters to maximize the quality of the data. First, we discarded individuals with a mean depth of coverage (DP) across sites lower than 8X. Second, given that the DP was heterogeneous across individuals, we treated all genotypes with DP<4 or with more than twice the individual mean DP as missing data to avoid genotype call errors due to low coverage or mapping errors. Third, we partitioned each scaffold into contiguous blocks of 1.5kb size, recording the number of SNPs and accessible sites for each block. To minimize regions with spurious SNPs (*e.g.*, due to mis-alignment or repetitive regions), we only kept blocks with more than 150bp of accessible sites and with a median distance among consecutive SNPs larger than 3bp. This resulted in a dataset consisting of 190 individuals with ~2.8Mb distributed across 11,770 blocks, with 284,287 SNPs and a total of 2,814,532 callable sites (corresponding to 2,530,245 monomorphic sites; Supplementary Table 7). Fourth, to minimize missing data, we only kept SNPs with at least 90% of called genotypes, and individuals with at least 75% of data across sites in the dataset. Finally, since many of the applied methods rely on the assumption of independence among SNPs, we generated a dataset by sampling one SNP per block, selecting the SNP with the lowest amount of missing data.

The PCA and the pairwise F_{ST} (estimated following Weir & Cockerham 1984) analyses were performed in R using the method implemented in the Bioconductor (release 3.4) package SNPrelate v1.8.0 (Zheng *et al.* 2012). We inferred the ancestry

proportions of all individuals based on *K* potential components using sNMF (Frichot *et al*. 2014) implemented in the R/Bioconductor release 3.4 package LEA v1.6.0 (Frichot & François 2015) with default settings. We examined *K* values from 1 to 12, and selected the *K* that minimized the cross entropy as suggested by Frichot *et al.* (2014). We performed the PCA, F_{ST} and sNMF analyses by applying an extra minor allele frequency (MAF) filter > (1/2*n*), where *n* is the number of individuals, such that singletons were discarded. To test for isolation by distance, we compared the estimated F_{ST} values to the pairwise geographic distances (measured as a straight line from the most northern sample, *i.e.,* Colome) between sample sites using a Mantel test, with significance assessed by 10,000 permutations of the genetic distance matrix.

Demographic Analyses

To uncover the colonization history of the Nebraska Sand Hills, we inferred the demographic history of populations based on the joint site frequency spectrum (SFS) obtained from the random anonymous genomic regions, using the composite likelihood method implemented in fastsimcoal2 v3.05 (Excoffier *et al.* 2013). In particular, we were interested in testing whether the Sand Hills populations were founded widely from across the ancestral range, or whether there was a single colonization event. We also aimed to quantify the current and past levels of gene flow among populations. We considered models with three populations corresponding to samples on the Sand Hills (*i.e.,* Ballard, Gudmundsen, SHGC, Arapaho, and Lemoyne) as well as off of the Sand Hills in the north (*i.e.,* Colome, Dogear, and Sharps) and in the south (*i.e.,* Ogallala, Grant, and Imperial). Specifically, we considered two alternative three-population demographic models to test whether the colonization of the Sand Hills most likely occurred from the north or from the south,

in a single event or serial events, thereby simultaneously quantifying the levels of gene flow between populations on and off of the Sand Hills (Supplementary Figure 2). In these models, we assumed that colonization dynamics were associated with founder events (*i.e.,* bottlenecks), and the number and place of which along the population trees were allowed to vary among models.

We constructed a three-dimensional (3D) SFS by pooling individuals from the three sampling regions (*i.e.,* on the Sand Hills as well as off of the Sand Hills in the north and south). For each of the 11,770 blocks of 1.5kb size, 30 individuals (*i.e.,* ten individuals from each of the three geographic regions) were selected such that all SNPs within a given block exhibited complete genotype information. Specifically, we selected the 30 individuals with the least missing data for each block, only including SNPs with complete genotype information across the chosen individuals, a procedure that maximized the number of SNPs without missing data while keeping the local pattern of linkage disequilibrium within each block. Note that we sampled genotypes of the same individuals within each block, but that at different blocks the sampled individuals might differ. For each block, we computed the number of accessible sites and discarded blocks without any SNPs.

The resulting 3D-SFS contained a total of 140,358 SNPs and 2,674,174 accessible sites (Supplementary Table 6). The number of monomorphic sites was based on the number of accessible sites, assuming that the proportion of SNPs lost with the extra DP filtering steps, not included in the mask file, was identical for polymorphic and monomorphic sites. Given that there was no closely related outgroup sequence available to determine the ancestral allele for SNPs within the random background regions, we analyzed the multidimensional folded site frequency spectrum, generated using Arlequin v.3.5.2.2 (Excoffier & Lischer 2010). For each

model, we estimated the parameters that maximized its likelihood by performing 50 optimization cycles (-L 50), approximating the expected SFS based on 350,000 coalescent simulations (-N 350,000), and using as a search range for each parameter the values reported in Supplementary Table 8.

The model best fitting the data was selected using Akaike's information criterion (AIC; Akaike 1974). Because our dataset likely contains linked sites, the confidence for a given model is likely to be inflated. Thus, to compare models, we generated bootstrap replicates with one SNP sampled from each 1.5kb block, which we assumed to be independent. We estimated the likelihood of each bootstrap replicate based on the expected SFS obtained for each model with the full set of SNPs (following Bagley *et al*. 2017). Furthermore, we obtained confidence intervals for the parameter estimates by a non-parametric block-bootstrap approach. To account for linkage disequilibrium, we generated 100 bootstrap datasets by sampling the 11,770 blocks for each bootstrap dataset with replacement in order to match the original dataset size. For each dataset, we performed two independent runs using the parameters that maximized the likelihood as starting values. We then computed the 95 percentile confidence intervals of the parameters using the R-package boot v.1.3-18.

Selection Analyses

We used VCFtools v.0.1.12b (Danecek *et al.* 2011) to calculate Weir and Cockerham's F_{ST} (1984) in the *Agouti* region. We pooled the 11 sampling locations into three populations (*i.e.,* one population on the Sand Hills and two populations off of the Sand Hills: one in the north and one in the south, see "Demographic Analyses") and calculated F_{ST} in a sliding window (window size 1kb, step size 100bp) as well as

on a per-SNP basis within *Agouti* to identify highly differentiated candidate regions for local adaptation.

To control for potential hierarchical population structure as well as past fluctuations in population size, we also measured genetic differentiation using the HapFLK method (Fariello *et al.* 2013). HapFLK calculates a global measure of differentiation for each SNP (FLK) or inferred haplotype (HapFLK) after having rescaled allele/haplotype frequencies using a population kinship matrix. We calculated the kinship matrix using the complete dataset, excluding the scaffolds containing *Agouti.* We launched the HapFLK software using 40 independent runs (--nfit 40) and –K 40, only keeping alleles with a MAF > 0.05. We obtained the neutral distribution of the HapFLK statistic by running the software on 1,000 neutral simulated datasets under our best demographic model.

To map potential complete selective sweeps in the *Agouti* region, we utilized the CLR method (Nielsen *et al.* 2005) as implemented in the software Sweepfinder2 (DeGiorgio *et al.* 2016). For the analysis, we used the *P. maniculatus rufinus Agouti* sequence to identify ancestral and derived allelic states (see "Variant Calling and Filtering"). We ran Sweepfinder2 with the "–su" option, defining grid-points at every variant and using a pre-computed background SFS. We calculated the cutoff value for the CLR statistic using a parametric bootstrap approach (as proposed by Nielsen *et al.* (2005)). For this purpose, we re-ran the CLR analysis on 10,000 datasets simulated under our inferred neutral demographic model for the Sand Hills populations in order to reduce false-positive rates (see Crisci *et al*. 2013; Poh *et al*. 2014).

ACKNOWLEDGEMENTS

We thank J. Larson and K. Turner for laboratory assistance; E. Kay, E. Kingsley, and M. Manceau for field assistance; J. Demboski and the Denver Museum of Nature and Science for logistical support; the University of Nebraska-Lincoln for use of facilities and/or permission to collect mice at Cedar Point Biological Station, Gudmundsen Sandhills Laboratory, and Arapaho Prairie; and J. Chupasko for curation assistance.

REFERENCES

Akaike H (1974) New look at statistical-model identification. IEEE Transactions on Automatic Control 19: 716–723.

Aeschbacher S, Burger R (2014). The effect of linkage on establishment and survival of locally beneficial mutations. Genetics 197: 317-336.

Bagley RK, Sousa VC, Niemiller ML, Linnen CR (2017). History, geography, and host use shape genome-wide patterns of genetic variation in the redhead pine sawfly. Molecular Ecology 26: 1022-1044.

Barsh GS (1996). The genetics of pigmentation: from fancy genes to complex traits. Trends Genet 12: 299–305.

Browning SR, Browning BL (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097.

Bulmer MG (1972). The genetic variability of polygenic characters under optimizing selection, mutation and drift. Genet Res 19: 17-25.

Bultman SJ, Klebig ML, Michaud EJ, Sweet HO, Davisson MT, Woychik RP (1994). Molecular analysis of reverse mutations from nonagouti (a) to black-and-tan (a (t)) and white-bellied agouti (Aw) reveals alternative forms of agouti transcripts. Genes & Development 8: 481-490.

Chaves JA, Cooper EA, Hendry AP, Podos J, De Leon LF, Raeymaekers JA, et al. (2016). Genomic variation at the tips of the adaptive radiation of Darwin's finches. Molecular Ecology 25: 5282-5295.

Clarke B, Murray J (1962). Changes of gene-frequency in *Cepaea nemoralis*: the estimation of selective values. Heredity 17: 467-476.

Comeault AA, Soria-Carrasco V, Gompert Z, Farkas TE, Buerkle CA, Parchman TL, et al. (2014). Genome-wide association mapping of phenotypic traits subject to a range of intensities of natural selection in *Timema cristinae*. Am Nat 183: 711–727.

Cook LM, Mani GS (1980). A migration-selection model for the morph frequency variation in the peppered moth over England and Wales. Biol J Linn Soc 13: 179–198.

Crisci J, Poh Y-P, Bean A, Simkin A, Jensen JD (2012). Recent progress in polymorphism-based population genetic inference. J Heredity 103: 287-296.

Crisci J, Poh Y-P, Mahajan S, Jensen JD (2013). On the impact of equilibrium assumptions on tests of selection Front Genet 4: 235.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The Variant Call Format and VCFtools. Bioinformatics 27: 2156-2158.
DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R (2016). SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics 32:1895-7.
DePristo M, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43: 491-498.
Dice LR (1941). Variation of the deer-mouse (*Peromyscus maniculatus*) on the Sand Hills of Nebraska and adjacent areas. Contrib. Lab. Vertebrate Biol. Univ. Michigan 15: 1-19.
Dice LR (1947). Effectiveness of selection by owls of deer mice (*Peromyscus maniculatus*) which contrast in color with their background. Contrib. Lab. Vertebrate Biol. Univ. Michigan 34: 1-20.
Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013). Robust demographic inference from genomic and SNP data. PLoS Genet 9 :e1003905.
Excoffier L, Lischer HE (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10: 564-567.
Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193: 929-941.
Frichot E, François O (2015). LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6: 925-929.
Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014). Fast and efficient estimation of individual ancestry coefficients. Genetics 196: 973-983.
Gompert Z, Lucas LK, Nice CC, Buerkle CA (2013). Genome divergence and the genetic architecture of barriers to gene flow between *Lycaeides* and *L. melissa*. Evolution 67: 2498–2514.
Haldane JBS (1930). *Enzymes*. London, New York, Longmans, Green and Co.
Haldane JBS (1957). The cost of natural selection. J Genet 55: 411-524.
Hoekstra HE, Drumm KE, Nachman MW (2004). Ecological genetics of adaptive color polymorphism in pocket mice: Geographic variation in selected and neutral genes. Evolution 58: 1329–1341.
Jackson IJ (1994). Molecular and development genetics of mouse coat color. Annu Rev Genet 28: 189-217.
Jensen JD, Foll M, Bernatchez L (2016). Introduction: the past, present, and future of genomic scans for selection. Molecular Ecology 25: 1-4.
Joost S, Vuilleumier S, Jensen JD, Schoville S, Leempoel K, Stucki S, *et al*. (2013). Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. Molecular Ecology 22: 3659-3665.
Kingsley SP, Manceau M, Wiley CD, Hoekstra HE (2009). Melanism in *Peromyscus* is caused by independent mutations in *Agouti*. PLoS One 4:e6435.
Kirkpatrick M, Barton N (2006). Chromosome inversions, local adaptation, and speciation. Genetics 173: 419-434.
Lenormand T. (2002). Gene flow and the limits to natural selection. Cell 17: 183-189.
Lenormand T, Otto SP (2000). The evolution of recombination in a heterogeneous environment. Genetics 156: 423-438.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25:2078-9.

Li H (2011). Improving SNP discovery by base alignment quality. Bioinformatics 27: 1157–1158.
Linnen CR, Hoekstra HE (2010). Measuring natural selection on genotypes and phenotypes in the wild. Cold Spring Harbor Symposia on Quantitative Biology 74: 155-68.
Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009). On the origin and spread of an adaptive allele in *Peromyscus* mice. Science 325: 1095-1098.
Linnen CR, Poh Y-P, Peterson BK, Barrett RDH, Larson JG, Jensen JD, et al. (2013). Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339: 1312-1316.
Lunter G, Goodson M (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research 21: 936-939.
Mallarino R, Linden TA, Linnen CR, Hoekstra HE (2017). The role of isoforms in the evolution of cryptic coloration in *Peromyscus* mice. Molecular Ecology 26: 245-258.
Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE (2010). Convergence in pigmentation at multiple levels: mutations, genes and function. Phil Trans R Soc Lond B Biol Sci 365: 2439-2450.
Manceau M, Domingues VS, Mallarino R, Hoekstra HE (2011). The developmental role of *Agouti* in color pattern evolution. Science 331: 1062-1065.
Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10-12.
Maynard Smith J (1977). Why does the genome not congeal. Nature 268: 693-696.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al*. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297-303.
Mills MG, Patterson LB (1994). Not just black and white: pigment pattern development and evolution in vertebrates. Semin Cell Biol 20: 72-81.
Montgomerie R (2006). Analyzing colours – In Bird Coloration, Volume I: Mechanisms and Measurements. (eds. G. E. Hill, K. J. McGraw) Harvard University Press, Cambridge, MA), pp. 90–147.
Montgomerie R (2008). CLR, version 1.05.
Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005). Genomic scans for selective sweeps using SNP data. Genome Research 15: 1566-1575.
Orr HA (1998). The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52: 935-949.
Pfeifer SP (2017). From next-generation resequencing reads to a high quality variant dataset. Heredity 118: 111-24.
Pickrell JK, Pritchard JK (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8: e1002967.
Poh Y-P, Domingues V, Hoekstra HE, Jensen JD (2014). On the prospect of identifying adaptive loci in recently bottlenecked populations: a case study in beach mice. PLoS One 9: e110579.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
Ross KG and Keller L (1995). Joint influence of gene flow and selection on a reproductively important genetic polymorphism in the fire ant *Solenopsis invicta*. Am Nat 146: 325-248.

Rousset F (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145: 1219-1228.
Sandoval CP (1994). The effects of the relative geographic scales of gene flow and selection on morph frequencies in the walkingstick *Timema cristinae*. Evolution 48: 1866-1879.
Smit AFA, Hubley R, Green P (2013-2015). RepeatMasker Open-4.0 <http://www.repeatmasker.org>.
Smouse PE, Long JC, Sokal RR (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Systematic Zoology 35: 627-632.
Sokal RR (1979). Testing statistical significance of geographic variation patterns. Systematic Zoology 28: 227-232.
Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, Halldorsdottir SS, et al. (2004). A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. Proc Natl Acad Sci USA 101: 4712-4717.
Tabachnick B, Fidell LS (2001). Using Multivariate Statistics. Allyn and Bacon, Boston, ed. 4t.
Tigano A, Friesen VL (2016). Genomics of local adaptation with gene flow. *Molecular Ecology* 25: 2144-2164.
Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine, et al. (2013). From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. Current Protocols 43: 11.10.1-11.10.33.
Vrieling H, Duhl DM, Millar SE, Miller KA, Barsh GS (1994). Differences in dorsal and ventral pigmentation result from regional expression of the mouse agouti gene. Proc Natl Acad Sci USA 91: 5667-5671.
Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.
Yeaman S, Otto SP (2011). Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. Evolution 65: 2123-2129.
Yeaman S, Whitlock MC (2011). The genetic architecture of adaptation under migration-selection balance. Evolution 65: 1897-1911.
Zheng X, Levin D, Shen J, Gogarten SM, Laurie C, Weir BS (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 15: 3326-3328.
Zhou X, Carbonetto P, Stephens M (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics 9: e1003264.

FIGURES

Figure 1. Sampling locations on (light brown) and off (dark brown) of the Nebraska Sand Hills. Sampling spanned a 330 km transect beginning ~120 km south of the Sand Hills and ending ~120 km north of the Sand Hills.

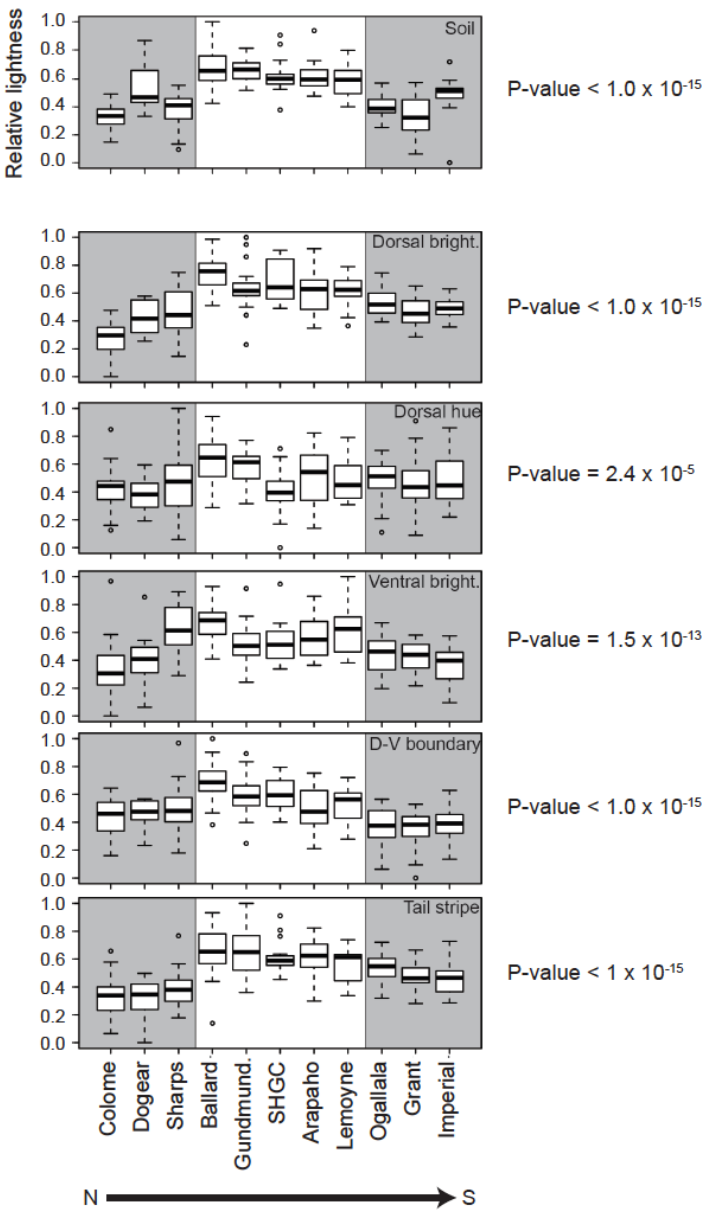

Figure 2. Soil color and mouse color traits are lighter on the Nebraska Sand Hills. Box plots for each sampling site were produced using scaled (range: 0-1), normal-quantile transformed values for: soil brightness (B2), dorsal brightness (PC1), dorsal hue (PC2), ventral brightness (PC3), dorsal-ventral (d-v) boundary, and tail stripe. Soil is shown on the separated top panel, and the five pigment traits are given below. In all plots, higher values correspond to lighter/brighter color (y-axis). Grey shading indicates off the Sand Hills sites. Soil and five mouse color traits were significantly lighter on the Sand Hills (dorsal brightness: $F_1 = 218.6$, $P < 1 \times 10^{-15}$; dorsal hue: $F_1 = 18.4$, $P = 2.4 \times 10^{-5}$; ventral brightness: $F_1 = 61.0$, $P = 1.5 \times 10^{-13}$; dorsal-ventral boundary: $F_1 = 104.6$, $P < 1 \times 10^{-15}$; tail stripe: $F_1 = 170.2$, $P < 1 \times 10^{-15}$), with the corresponding *P*-values given to the right of each panel. A sixth color trait [ventral hue (PC4); ($F_1 = 1.1$, $P = 0.30$)] and four non-color traits (total body length ($F_1 = 0.11$, $P = 0.74$), tail length ($F_1 = 0.18$, $P = 0.67$), hind foot length ($F_1 = 3.0$, $P = 0.084$), and ear length ($F_1 = 0.48$, $P = 0.49$)) did not differ significantly between habitats (data not shown; see text).

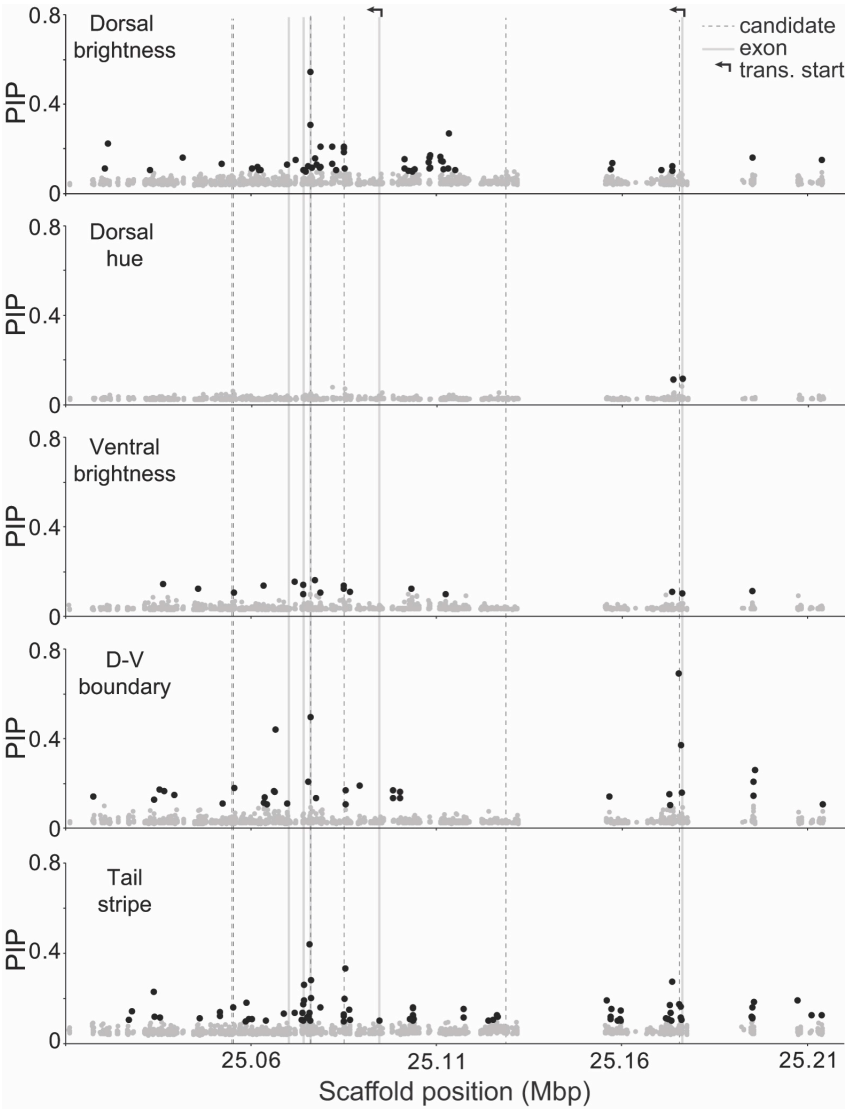

Figure 3. Association mapping results for *Agouti* SNPs and five color traits. Each point represents one of 2,148 *Agouti* SNPs included in the association mapping analysis, with position (on scaffold 16, see Methods) indicated on the x-axis and posterior inclusion probabilities (PIP) estimated using a BSLMM in GEMMA on the y-axis. PIP values, which approximate the strength of the association between genotype and phenotype, were averaged across 10 independent GEMMA runs. Black dots indicate SNPs with PIP > 0.1, dashed lines give the location of six candidate SNPs identified in a previous association mapping analysis of a single population (Linnen *et al.* 2013), grey lines give the location of *Agouti* exons, and arrows indicate two alternative transcription start sites.

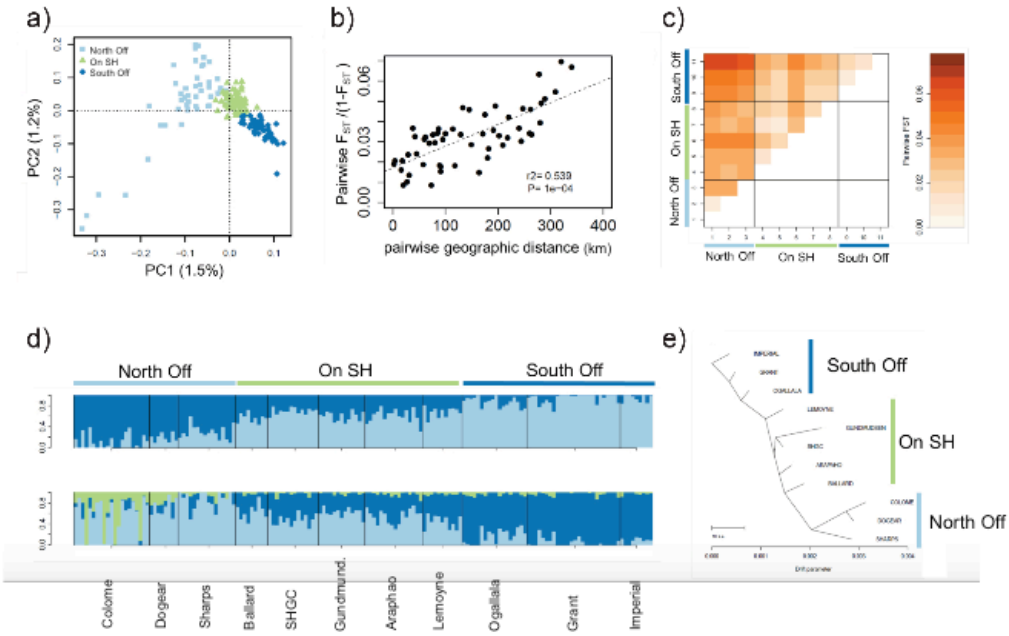

Figure 4. Genetic structure of populations on and off the Sand Hills. (**a**) Principal Component Analysis (PCA) shows that individuals cluster according to their geographic sampling location, with a clear separation of individuals sampled on the Sand Hills (SH; green) from off the Sand Hills (blue). (**b**) An isolation-by-distance model is supported by a significant correlation between pairwise genetic differentiation (F_{ST}/(1-F_{ST})) and pairwise geographic distance among sampling sites. Significance of the Mantel test was assessed through 10,000 permutations. (**c**) Pairwise F_{ST} values among samples. (**d**) Individual admixture proportions (inferred using *sNMF*) are consistent with isolation by distance and reduced differentiation among populations. The number of clusters (*K*) best explaining the data was assessed as the *K* value reaching the lowest minimal cross-entropy at *K*=3, but with similar distribution of values for *K* between 1 and 3 (*K*=2 and *K*=3 shown here). (**e**) TreeMix results for the tree that best fit the covariance in allele frequencies across samples. All analyses are based on a subset of 161 individuals showing a mean depth of coverage larger than 8X, with all genotypes possessing a minimum coverage of 4X, sampling one SNP for each 1.5kb block, resulting in a data set with 5,412 SNPs (Supplementary Table 6).

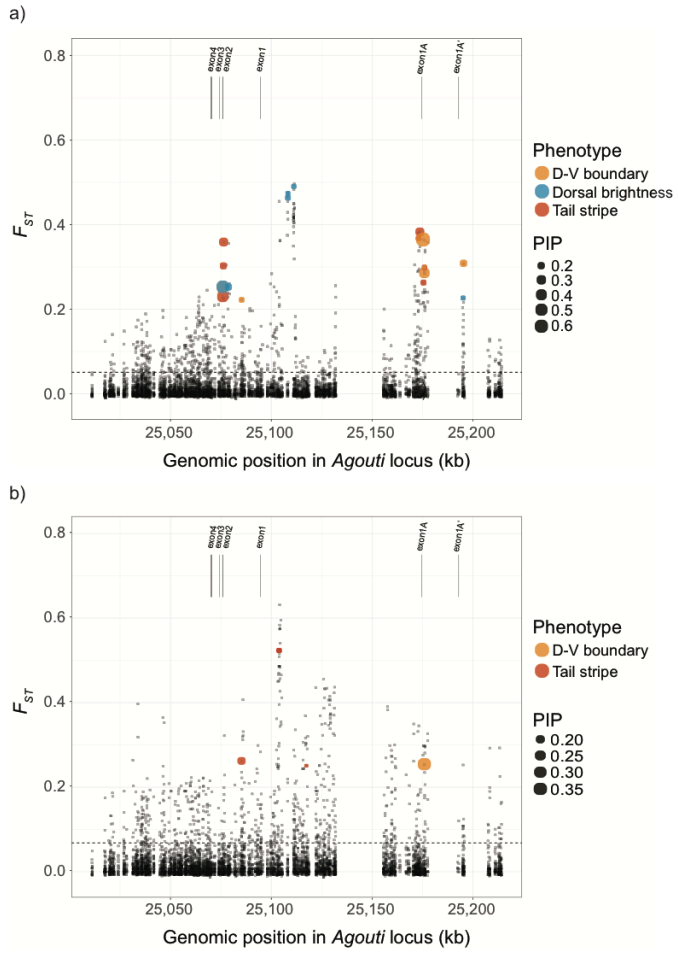

Figure 5. **Genetic differentiation across the *Agouti* locus**. Genetic differentiation (F_{ST}, calculated in 1kb windows with a 100bp step size) observed between (**a**) populations on the Sand Hills and populations south of the Sand Hills, and (**b**) populations north and south of the Sand Hills. Dotted lines indicate the genome-wide average F_{ST} between these comparisons. As shown, differentiation is generally greatly elevated across the region, with a number of highly differentiated SNPs particularly when comparing on vs. off Sand Hills populations (panel a), which correspond to SNPs associated to different aspects of the cryptic phenotype (PIP scores are indicated by the size of the dots as shown in the legend, and significant SNPs are labeled with respect to the associated phenotype). Populations to the north and south of the Sand Hills (panel b) are also highly differentiated in this region, but with many fewer SNPs associated with the cryptic phenotype.

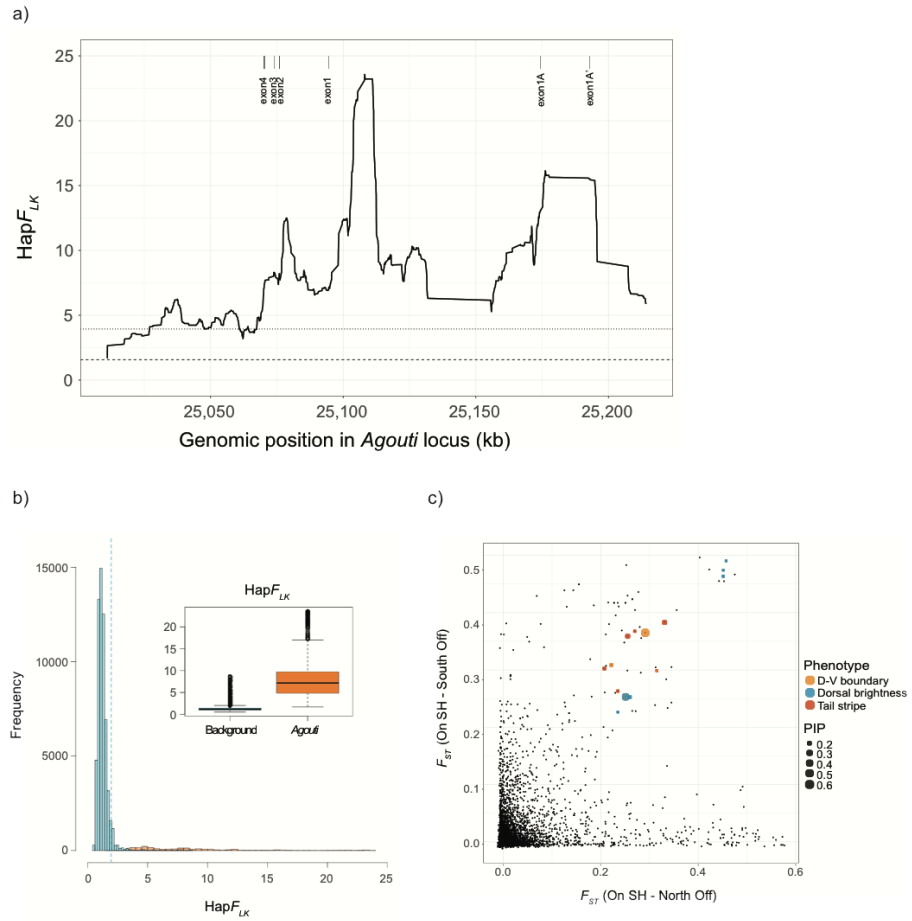

Figure 6. Selection analyses. (**a**) HapFLK profile for the *Agouti* scaffold observed between populations on and off the Sand Hills. Dotted and dashed horizontal lines represent the significance thresholds for the HapFLK statistic using neutral simulations and background data, respectively. (**b**) Distributions of HapFLK values for background (blue) and *Agouti* (orange). The background values are calculated from the genome-wide random background SNPs. Boxplots (insert) are an alternative representation of the two distributions depicted in the main graph. The dashed line indicates the significance threshold using parametric bootstrapping of the best demographic model. (**c**) Genetic differentiation per SNP in *Agouti*. The x-axis represents values obtained from comparing Sand Hills populations to those off of the Sand Hills (north), while the y-axis compares Sand Hills populations to those off of the Sand Hills (south). SNPs associated with phenotypic traits are indicated with their PIP value. Only candidate SNPs with a F_{ST} value larger than 0.2 in both populations and PIP values larger than 0.15 are included.