Dispersal inference from population genetic variation using a convolutional neural network

Chris C. R. Smith[1], Silas Tittes[1], Peter L. Ralph[1], and Andrew D. Kern[1]

[1]University of Oregon Institute of Ecology and Evolution

August 25, 2022

Abstract

The geographic nature of biological dispersal shapes patterns of genetic variation over landscapes, so that it is possible to infer properties of dispersal from genetic variation data. Here we present an inference tool that uses geographically-referenced genotype data in combination with a convolutional neural network to estimate a critical population parameter: the mean per-generation dispersal distance. Using extensive simulation, we show that our deep learning approach is competitive with or outperforms state-of-the-art methods, particularly at small sample sizes (e.g., n=10). In addition, we evaluate varying nuisance parameters during training—including population density, population size changes, habitat size, and the size of the sampling window relative to the full habitat—and show that this strategy is effective for estimating dispersal distance when other model parameters are unknown. Whereas competing methods depend on information about local population density or accurate identification of identity-by-descent tracts as input, our method uses only single-nucleotide-polymorphism data and the spatial scale of sampling as input. These features make our method, which we call `disperseNN`, a potentially valuable new tool for estimating dispersal distance in non-model systems with whole genome data or reduced representation data. We apply `disperseNN` to 12 different species with publicly available data, yielding reasonable estimates for most species. Importantly, our method estimated consistently larger dispersal distances than mark-recapture calculations in the same species, which may be due to the limited geographic sampling area covered by some mark-recapture studies. Thus genetic tools like ours complement direct methods for improving our understanding of dispersal.

Introduction

Organisms vary greatly in their capacity to disperse across geographic space. Indeed, the movement of individuals or of gametes across a landscape, in part, determines the spatial scale of genetic differentiation and the spread of adaptive variants across natural populations (Broquet and Petit, 2009). Consequently, understanding dispersal is relevant for conservation biology (Driscoll et al., 2014), studying climate change response and adaptation (Travis et al., 2013), managing invasive and disease vector populations (Harris et al., 2009; Orsborne et al., 2019), phylogeography (Kadereit et al., 2005), hybrid zones and speciation (Barton, 1979), microbial community ecology (Evans et al., 2017), and for parameterizing models in ecology and evolution (Barton et al., 2002). Despite the importance of dispersal, it remains challenging to obtain estimates for dispersal distance in many species.

Some methods infer dispersal distance by directly observing individual movement, using radio-tracking technology, or by tagging and recapturing individuals in the field. However, such measurements can be expensive to obtain and lead to estimates with high uncertainty. Furthermore, they do not always provide a complete picture of the *effective* dispersal rate—that is, how far successfully-reproducing individuals travel from their birth location on average over many generations (Bradburd and Ralph, 2019). This long term average is often the quantity of interest as it is more relevant for understanding population structure, evolutionary dynamics of selected alleles, and long-term changes to a species' range.

Another type of method infers (effective) dispersal distance from a single temporal sample, without directly observing movement of individuals. Such inference is possible because population genetics theory predicts how demographic parameters such as the rate of gene flow across the landscape affect the genetic variation of a population (Barton et al., 2013). To infer dispersal distance, current population-genetics-based estimators (Rousset, 1997; Ringbauer et al., 2017) use geographically-referenced DNA sequences and can obtain useful estimates of the per-generation dispersal distance, without the need for tracking or recapturing individuals.

Importantly, current population-genetics-based estimators require additional data that can be prohibitively expensive, especially for non-model species: either an independent estimate of population density (Rousset, 1997), or genomic identity-by-descent blocks (Ringbauer et al., 2017). Specifically, the seminal method of Rousset (1997) is designed for estimating neighborhood size, N_{loc}, which can be thought of as the number of neighboring individuals or potential mates that are within a few multiples of the dispersal distance (Wright, 1946). Wright defined neighborhood size as $N_{loc} = 4\pi D\sigma^2$, where σ is the dispersal distance and D is the population density. Therefore the accuracy of Rousset's method depends on having a good *a priori* estimate of population density. One way to jointly infer dispersal and density works by modeling genomic identity-by-descent tracts (e.g., Barton et al., 2013; Baharian et al., 2016; Ringbauer et al., 2017). Similarly, the program MAPS (Al-Asadi et al., 2019) uses identity-by-descent information to infer heterogeneous dispersal and density across a landscape. Although powerful when applied to high quality data, these methods are limited by the availability of confident identity-by-descent blocks; this type of data remains unavailable or difficult to estimate for many species. Thus for most study species we are stopped short of quantifying dispersal distance from population genetic data.

Another type of population-genetics-based method estimates *relative* migration rates, for example EEMS (Petkova et al., 2016), FEEMS (Marcus et al., 2021), and other landscape genetics tools. Although such methods work well for some applications, such as identifying barriers to dispersal, they don't inform us about the magnitude of dispersal, e.g., meters per generation. Furthermore, these and related tools model gene flow using an approximate analogy to electrical resistance which can produce misleading results especially in the presence of biased migration (Lundgren and Ralph, 2019). In the current paper we set out to develop a method for estimating dispersal distance that can be applied widely, including in non-model species without good assemblies or knowledge of population density.

To do this we use simulation-based inference via deep learning to infer dispersal from genotype data directly. Deep learning is a form of supervised machine learning that builds a complex function between input and output involving successive layers of transformations through a "deep" neural network. An important advantage of this class of methods is their ability to handle many correlated input variables without knowledge of the variables' joint probability distribution. Like all supervised machine learning methods, deep neural networks can be trained on simulated data, which bypasses the need to obtain empirical data for training (Schrider and Kern, 2018). Over the past few years, deep learning has been used in a number of contexts in population genetics: for example, inferring demographic history in *Drosophila* (Sheehan and Song, 2016), detection of selective sweeps (Kern and Schrider, 2018), detecting adaptive introgression in humans (Gower et al., 2021), identifying geographic origin of an individual using their DNA (Battey et al., 2020a), and estimating other population genetic parameters like recombination rate (Flagel et al., 2019).

We present the first use of deep learning for estimation of spatial population genetic parameters. Our method, called `disperseNN`, uses forward in time spatial genetic simulations (Haller and Messer, 2019; Battey et al., 2020b) to train a deep neural network to infer the mean, per-generation dispersal distance, from a single population sample of single nucleotide polymorphism (SNP) genotypes, e.g., whole genome data or RADseq data. We show that `disperseNN` is more accurate than two competing methods (Rousset, 1997; Ringbauer et al., 2017) at inferring dispersal distance, particularly for small to moderate sample sizes, or when identity-by-descent tracts cannot be reliably inferred. After exploring potential shortcomings of our method, we demonstrate its utility on several empirical datasets from a broad range of taxa. The `disperseNN` software is available from https://github.com/kr-colab/disperseNN, where we have also provided a pre-trained model for ease of prediction in new systems.

Results

Dispersal estimation using deep neural networks

We use a convolutional neural network (CNN) trained on simulated data to infer the average per-generation dispersal distance (Figure 1). The CNN takes two pieces of data as input: (1) a genotype matrix, and (2) the width of the geographic sampling area. The genotype matrix is put through the network's convolution layers, while the geographic sampling width is used downstream and is important for conveying the physical scale of sampling. The output from the CNN is a single estimate of the dispersal parameter, σ. Our software package, `disperseNN`, has several inference-related functionalities: (i) training the CNN on simulated data, (ii) predicting σ using simulated or empirical data, and (iii) pre-processing steps for empirical data. In addition, the `disperseNN` package includes a network pre-trained by us that can be used to estimate dispersal distance from empirical data without additional training.

The training data for `disperseNN` are simulated using a continuous-space SLiM model following Battey et al. (2020b). In this model, each offspring disperses from their parent's location an independent bivariate Gaussian displacement with mean zero and standard deviation σ in each direction. We refer to σ as "the dispersal parameter", although the mean distance dispersed is roughly 1.26σ. In addition to dispersal, σ also determines the mating and competitive interaction distances in our simulation model. `disperseNN` provides an estimate of σ in the same units as its second input, the width of the sampling area, from training.

Training with `disperseNN` consists of: deciding on training distributions for σ and other parameters, using a spatial model to simulate training data, and handing the simulation output and targets (true σ) to `disperseNN` for training the CNN. The analysis pipeline for predicting on simulated data is similar to that of training, while predicting on empirical data involves basic pre-processing of the input data before using `disperseNN` to estimate σ. Below, we present findings from several experiments using `disperseNN`, each with its own set of parameters for simulation and training. We describe each experiment briefly in the Results section, and reference different sets of parameters that correspond to each experiment, e.g, "Parameter Set 1", "Parameter Set 2", etc. Full details about the different parameter sets are in the Materials and Methods section.

Comparison with existing methods

We evaluated the accuracy of our method on simulated datasets with a range of σ values (Parameter Set 1), using the relative absolute error (RAE) to measure prediction accuracy for each estimate:

$$\text{RAE} = \left| \frac{\text{estimated } \sigma - \text{true } \sigma}{\text{true } \sigma} \right| \tag{1}$$

For comparing accuracy between training runs or between methods, we calculate the mean relative absolute error (MRAE) averaged across all test datasets. We found `disperseNN` estimates σ more accurately than previous genetics-based methods (Figure 2). At small sample sizes ($n = 10$), `disperseNN` was dramatically more accurate than both the Rousset (1997) method and the program from Ringbauer et al. (2017) called `IBD-Analysis` (MRAE=0.12, 0.39, and 26.24, respectively). Furthermore, the Rousset method and `IBD-Analysis` produced undefined output and convergence errors for 16.4% and 4.6% of test datasets, respectively. For Rousset's method, at least, this is due to a negative slope in the least squares fit of genetic distance versus geographic distance, which happens more frequently with a small sample size.

`disperseNN` remained the most accurate method when the sample size was large ($n = 100$) in part due to a bias using `IBD-Analysis` (MRAE = 0.08, 0.19, and 0.36, respectively). Estimates from `IBD-Analysis` were consistently, slightly overestimated for $n = 100$, and estimates from the Rousset method were underestimated on average, likely due to model misspecification. In particular, Ringbauer et al. (2017) reported moderate overestimation of σ when sampling locations were sampled uniformly, instead of regularly spaced in a grid. The sample locations in our analysis are random and irregular (Figure S1), which likely accounts for the bias using `IBD-Analysis`. The `IBD-Analysis` program may perform best on data with regular spacing between sample locations when $n = 100$, however the bias in this case results from including biological realism. It is also important to note that we provided `IBD-Analysis` with true identity-by-descent tracts, when in reality perfectly inferred identity-by-descent tracts are not available for most species, and inferring identity-by-descent tracts from (perfectly phased) SNPs reduces the accuracy of `IBD-Analysis` (Figure S2).

Larger numbers of SNPs further improved the accuracy of `disperseNN`, although with diminishing returns (Figures S3, S4). Larger values for σ showed correspondingly larger errors (Figure 2), however *relative* error was nearly constant across the range of true σ (Figure S5). In addition, `disperseNN` and the Rousset method slightly underestimated σ when the true value approached the maximum of the examined range. This occurs because there is little spatial structure when σ is large relative to the habitat width. This observation from simulated data suggests we might expect `disperseNN` to have limitations when analyzing populations with very little spatial structure caused by isolation-by-distance.

Varying individual nuisance parameters during training

A common concern with supervised machine learning methods is that data used for prediction may fall outside of the training distribution. If the training set was simulated with, for example, a small population density, should we expect the trained network to accurately estimate σ if the test data have a large density? We set out to explore limitations of `disperseNN` using deliberately misspecified simulations, including out-of-sample (i) population density, (ii) ancestral population size, (iii) habitat size, and (iv) restricted sampling area relative to the full habitat. We individually address each scenario by augmenting the training set, which ultimately allows us to circumvent the problem of each nuisance parameter being unknown. This procedure is explained in more detail below.

First, we obtained a baseline accuracy-level for a "naive" model by training `disperseNN` on data where all simulation parameters were fixed except for σ (Parameter Set 2). This resulted in an MRAE of 0.12 using test data with all parameters drawn from the same distribution as the training set. We next used the model trained on Parameter Set 2 to estimate σ in test data where one of the aforementioned nuisance parameters is misspecified to varying degrees, i.e., drawn from outside the range of the training set (Parameter Sets 3, 5, 7, 9). Such model misspecification reduced the accuracy of σ estimation (Figure 3, column 2). This reduction in accuracy was most pronounced for misspecified population density and habitat width parameters (MRAE = 0.36 for each). The other scenarios also increased error, although more moderately. When a fixed habitat width was assumed, 23% of predictions were larger than the maximum σ from training; for other nuisance parameters all predictions fell within the range of σ used in training.

Having observed the effect of misspecification due to assuming particular values for nuisance parameters, we next assessed a training strategy for dealing with each unknown parameter. For each misspecification scenario, we assign a distribution to the unknown model parameter and allow the parameter to vary across training simulations (reusing Parameter Sets 3, 5, 7, 9). Using the new training set, `disperseNN` learned to accurately estimate σ when individual nuisance parameters were unknown, with error levels approaching the original MRAE (Figure 3, column 3; Parameter Sets 3, 5, 7, 9). To reiterate, this procedure varied a single unknown parameter at a time, not in combination. Essentially by treating each unknown parameter as a nuisance parameter during training, the model can become agnostic to the unknown parameter—or else learn a representation for the parameter such that σ can be calculated conditional on the learned parameter. This ability is critical for applying supervised learning methods for estimating σ where model parameters other than σ are unknown.

Although `disperseNN` was able to predict σ after including variation in each nuisance parameter in the training set, we next show that extrapolation is limited in some cases for unfamiliar parameter values, i.e., values outside of the distribution used for training. In the preceding trial the same distribution was used for both training and prediction. Next we assessed `disperseNN`'s ability to extrapolate at very large values of each nuisance parameter (Parameter Sets 4, 6, 8, 10), beyond the range used in training (Parameter Sets 3, 5, 7, 9). Results from this experiment were varied (Figure 3, rightmost column): predictions at out-of-sample values of density and ancestral population size were unreliable, but we were able to predict at large, out-of-sample habitat sizes and sampling areas quite well. Of note, using very large habitat sizes resulted in only a single estimate being 1% larger than the maximum σ from training.

Dealing with multiple nuisance parameters

After finding a successful training strategy for dealing with individual nuisance parameters, we next sought to train a network for general use in estimating σ where multiple parameters are unknown. The resulting network is what we refer to as "the pre-trained network". To do this, we used large ranges for parameters that

control: (i) dispersal distance, (ii) population density, (iii) ancestral population size, (iv) timing of population size change, (v) habitat size, and (vi) the size of the sampling area relative to the full habitat (Parameter Set 11). Furthermore, we exposed `disperseNN` to a range of different sample sizes between 10 and 100 by padding the genotype matrix out to 100 columns during training. Training simulations used 5,000 SNPs sampled from a single 100 megabase chromosome; this approach resembles a RADseq experiment, as the loci are spaced out on the chromosome and may be considered mostly unlinked. Last, we collapsed the diploid genotypes output by SLiM into unphased genotypes; 0s, 1s, and 2s; representing the count of the minor allele at each variable site. Through validation with held-out, simulated data, we found that the final model was accurate across a wide range of nuisance parameter values, and showed roughly order-of-magnitude accuracy (MRAE=0.55; Figure 4).

We provide the learned weights and biases from the above pre-trained network for download as part of the `disperseNN` package. The pre-trained network can be used to quickly estimate σ from various species or simulated datasets without additional training or simulations. We note that the pre-trained network for `disperseNN` could in addition be an excellent starting place for transfer learning (Weiss et al., 2016) for specific organisms, sampling designs, or perhaps alternative datatypes (e.g., microsatellite mutations). Benchmarking the pre-trained model on our system, it took 6.5 seconds to estimate σ using a dataset of 10 individuals and 5,000 variants, with the majority of computation time spent loading software libraries and pre-processing the genotype matrix. While `disperseNN` can be trained with any number of SNPs, m, the pre-trained network uses $m = 5,000$. Therefore, if fewer than 5,000 variants are available, as in some RADseq datasets, then a new network must be trained to match the empirical number.

The pre-trained model will be more appropriate for some datasets than others. First, the model was trained on 10 to 100 individuals sampled across a region of known width. Therefore, data collected from a single location are not expected to give accurate predictions (unless the breeding locations were known and spatially distributed). In fact, we strictly avoid repeated sampling localities and ensure that each location is represented by only one individual. The pre-trained model uses 5,000 SNPs; padding the input genotypes with zeros will not suffice in this case, as we did not train with zero-padding. Although we aimed to produce a pre-trained model that is widely applicable, there were parts of parameter space that were not represented during training. Specifically, many of the attempted simulations either resulted in population extinction, or could not be simulated due to computational constraints. These factors skewed the realized training distributions (Figure S6). Therefore, we expect this model to be most applicable for populations that fall solidly inside of the training distribution. For example, the model we provide was trained with σ, population densities, and sampling windows as large as 78 km, 994 individuals per km^2, and 944 km, respectively.

Additional training will be beneficial in some situations. If independent estimates for nuisance parameters or better-informed "prior" ranges are available, new training data may be tailored using the better-informed values. Species range maps with detailed geographic boundaries can be simulated with SLiM (since version 3.5), which in most cases will be superior to the square map we used. Importantly, if empirical parameters fall outside of the training distributions used for the pre-trained network, e.g., very large sampling area, then new training data will need to be generated that reflect the real data.

Quantifying uncertainty

In addition to helping us generate training data, simulation also allows us to quantify uncertainty through validating our models on held-out test datasets. Indeed, our reported values for MRAE give a sense of how much error to expect when applying the method to real data, in so far as the data resemble a typical draw from our test simulations. For example, in the above experiments that included one or zero nuisance parameters, the MRAE from in-sample tests was on the order of 0.12. Therefore, using a model with MRAE of 0.12 we might expect future predictions to be off from the true values by about 12%. However, if the real data are not well represented by the simulations, for example if the density of the analyzed population does not resemble that of the training simulations, then predictions might be less accurate, or biased.

Since we get distinct estimates for each subset of m SNPS, we might also assess uncertainty by looking at the range of variation among these estimates, i.e., through non-parametric bootstrapping. Each subsample of m SNPs from the same set of sampled individuals gives a different estimate of σ because of the varying genealogical histories that underlie different subsets of genomic loci, so the range of variation reflects the uncertainty arising from this genealogical noise. However, note that the bootstrapped estimates are not

independent, because they come from a single set of individuals. The `disperseNN` program provides a built-in functionality for performing this bootstrapping procedure, and will report the distribution of estimates across replicate draws of m SNPs (each draw is made without replacement from the complete set of available SNPs, but the replicates are drawn independently and so may overlap).

Although the distribution of these estimates should reflect uncertainty somehow, it is not immediately clear how to convert this into a formal quantification of uncertainty. This distribution of estimates is not a sample from a well-calibrated posterior distribution (nor should we expect it to be): in the test data for the pre-trained model (Figure 4), the true σ was covered by the middle 95% range from the bootstrap distribution for only 51% of simulated datasets. However, we can inflate the interval obtained by a scalar value such that our bootstrap interval is better calibrated. On our validation set for the pre-trained model this scalar value is 3.8, which leads to intervals that cover the true value for 95% of our test simulations. (If $\hat{\sigma}$ is the mean of the bootstrap estimates, and a and b are the 2.5% and 97.5% quantiles, respectively, then the resulting interval is from $\hat{\sigma} + 3.8(a - \hat{\sigma})$ to $\hat{\sigma} + 3.8(b - \hat{\sigma})$.) However, if this is to be a recipe for a well-calibrated credible (or, confidence) interval, then it needs to apply regardless of the situation: i.e., the magnitude of the error should be a roughly constant multiple of the range of the bootstrap estimates. Happily, this is the case: we found the error to be roughly a constant multiple of the width of the range of bootstrap estimates. (Concretely, if σ is the true value, $\hat{\sigma}$ is the estimated value, and w is the range of values from 100 bootstrap estimates, then $|\sigma - \hat{\sigma}|/w$ has no significant associations with any of the model parameters; see Figure S7.)

In summary, this suggests that the middle 95% interval of bootstrap estimates, inflated by a factor of 3.8, can stand in for a 95% credible interval for results obtained from our pre-trained neural network. Of course, since this is an empirically derived result, we do not expect the same inflation value to be appropriate for other networks or for datasets not well-represented by the simulations in the training set for our pre-trained model.

Empirical findings

We used `disperseNN` to estimate σ from a diverse set of organisms using preexisting empirical datasets that were available in repositories online. The pre-trained `disperseNN` model works with either whole genome sequencing or RADseq data, because the model was trained on mostly-unlinked SNPs distributed throughout the genome and genotypes were not phased during training. For some empirical datasets we analyzed a subset of sample localities in order to keep the sampling area less than 1,000 km; accordingly, we report sample sizes and sampling widths from the subsampled region, rather than the full dataset. For each dataset, `disperseNN` converts the SNP table to a genotype matrix, finds the width of the sampling area from the sample locations, and hands the two inputs to the pre-trained CNN described above. Additionally, we bootstrapped each SNP table to obtain 1,000 replicates of 5,000 random SNPs and predicted σ in each to obtain a distribution of estimates. Table 1 shows the mean and approximate 95% credible interval of σ estimates for each empirical dataset.

When available, we report previous dispersal estimates from the literature. Independent estimates came from a variety of methods including mark-recapture, tracking devices, and the Rousset method. Overall we find a correlation ($r^2 = 0.39$; $p = 0.03$) between our estimates and previous estimates using different methods. We might expect each of the analyzed empirical datasets to deviate from our training set in some way. To get a rough estimate of the "distance" between an empirical dataset and our training set we calculated five summary statistics—nucleotide diversity, Tajima's D, F_{IS} (an estimate of inbreeding), observed heterozygosity, and expected heterozygosity—and calculated the Mahalanobis distance between the centroid of the training distribution and each dataset, according to: $D^2 = (x - m)^T \cdot C^{-1} \cdot (x - m)$, where D^2 is the Mahalanobis distance squared, x is a vector of summary statistics from an empirical dataset, m are the means of each summary statistic in the training data, and C^{-1} is the inverse covariance matrix of the summary statistics calculated on the training data. Thus, smaller distances have summary statistics more similar to the training distribution, and distances larger than 40 fall outside of the training distribution (Figure S8).

Zosterops: Réunion grey white-eye and Réunion olive white-eye are endemic to the island of Réunion with approximate land area of 2500 km^2. These populations' restricted range make them ideal for analyzing with our pre-trained model. We analyzed the RADseq data from Gabrielli et al. (2020) including 41 individuals

Species	Common name	Region	σ	95% CI	Previous	N_{loc}	n	S	M. dist.
Zosterops borbonicus	Réunion grey white-eye	Réunion	4.06	(1.44, 11.29)	NA	295	41	62	4.59
Peromyscus leucopus	white-footed mouse	New York	0.63	(0.26, 1.36)	0.03-0.11	-231	12	38	8.15
Anopheles gambiae	African malaria mosquito	Cameroon	8.40	(1.63, 39.22)	0.04-0.5	52	29	278	9.62
Bombus bifarius	two-form bumble bee	Washington	12.04	(4.57, 30.44)	1.2-5	1,147	14	273	10.47
Bombus vosnesenskii	yellow-faced bumble bee	California	6.29	(0.99, 31.11)	1.2-5	3,944	18	169	11.83
Hippoglossus hippoglossus	Atlantic halibut	Canada	3.50	(0.58, 27.64)	NA	-5,546	11	193	14.59
Crassostrea virginica	eastern oyster	Canada	1.24	(0.59, 3.52)	21.9	1,435	13	187	19.69
Canis lupus	grey wolf	N. America	12.80	(1.93, 87.63)	98-147	35	13	721	25.42
Helianthus petiolaris	prairie sunflower	Kansas	0.82	(0.32, 2.87)	0.156	9	11	204	45.28
Zosterops olivaceus	Réunion olive white-eye	Réunion	0.86	(0.22, 3.56)	NA	2,392	10	50	45.97
Helianthus argophyllus	silverleaf sunflower	Texas	0.85	(0.31, 3.33)	0.156	57	30	307	86.49
Arabidopsis thaliana	thale cress	Spain	1.11	(0.23, 4.12)	0.001	84	84	80	198.25
Arabidopsis thaliana	thale cress	Sweden	0.36	(0.16, 0.76)	0.001	84	35	325	428.17

Table 1: Empirical results. The σ column is the mean from 1000 subsamples of 5,000 SNPs. "95% CI" is the credible interval obtained from bootstrapping. The "Previous" column shows previously published estimates for dispersal distance. N_{loc} is the neighborhood size using the Rousset calculation. In other columns, n is sample size, S is the width of the sampling area in kilometers, and "M. dist." is the Mahalonobis distance from the center of the training distribution with respect to five summary statistics: nucleotide diversity, Tajima's D, inbreeding coefficient, observed heterozygosity, and expected heterozygosity.

and 7,657 SNPs from *Z. borbonicus* and 10 individuals and 6,103 SNPs from *Z. olivaceus*. Our estimate for *Z. borbonicus* was 4.1 km, however the estimate in *Z. olivaceus* was smaller, 0.9 km. Although we are not aware of other dispersal estimates in these species, the data curated by Paradis et al. (1998) include natal dispersal estimates for 75 birds, and the smaller species, comparable in size to *Zosterops*, have dispersal distances in the range of 1-20 km. The mean estimate for *Z. borbonicus* falls within the range from Paradis et al., and the estimates for *Z. olivaceus* are close. While the data for both *Zosterops* species are similar, summary statistics in *Z. olivaceus* were further from the centroid of the training distribution.

Peromyscus leucopus: From the white footed mouse RADseq dataset of Munshi-South et al. (2016) we analyzed 12 individuals collected from the New York City metropolitan area, with 5,536 SNPs. We estimated dispersal distance to be 630 m. For comparison, Keane (1990) and Jacquot and Vessey (1995) measured natal dispersal in white footed mice in rural locations. They reported mean dispersal of 85-109 m in males and 25-88 m in females, which is smaller than our estimate. However, their estimates are likely constrained to some degree by the small study areas used for recapture. Indeed not all mice were recaptured in Jacquot and Vessey (1995), leaving open the possibility of long distance movements outside of the study area. For example, Murie and Murie (1931) documented travel distances greater than 1 km in *Peromyscus maniculatus*. Occasional long distance dispersal may help reconcile the difference between previous estimates and ours.

Anopheles gambiae: From the whole genome resequencing dataset from the *Anopheles gambiae* 1000 Genome Consortium (2021) we analyzed 29 individuals with 11 million SNPs. Our estimate in *A. gambiae* of 8.4 km is substantially larger than mark-recapture estimates. For comparison, Epopa et al. (2017) measured individual *A. coluzzii* dispersal distances between 40 to 549 m over seven days; however the geographic study region was restricted to a single village. It is unclear to what degree long-distance dispersal in mosquitos contributes to effective dispersal and gene flow. Remarkably, the recent study of Huestis et al. (2019) captured *A. gambiae* and other mosquito species 40 m to 290 m above the ground, suggesting a wind-borne dispersal mechanism. Assuming average wind speeds, Huestes et al. estimated that each year tens of thousands of *A. gambiae* individuals migrate 10s or 100s of km in the atmosphere of the studied region. These findings suggest that dispersal potential in this species is considerably larger than once thought. Significant long-range dispersal in *A. gambiae* is consistent with some predictions in the species, as there is little genetic differentiation across portions of the species range (e.g., West Africa), while at broader scales structure is appreciable (Anopheles gambiae 1000 Genome Consortium, 2017)

Bombus: From the dataset of Jackson et al. (2018) we examined RADseq data from two bumble bee species, *B. bifarius* and *B. vosnesenskii* with samples sizes of 14 and 18, and 8,073 and 6,725 SNPs, respectively. Our estimated dispersal distances were 12.0 km and 6.3 km for them in turn. These species are eusocial, thus our dispersal estimate should reflect the distance traveled by queens that start successful nests. Mark-recapture analyses have found a minimum distance traveled by queens in other *Bombus* species of 1.2

km (Carvell et al., 2017) and 3-5 km (Lepais et al., 2010). These estimates are particularly relevant, as they measure natal dispersal from the birth location of the queen. Even so, these values represent a lower bound distance that queens disperse, as their was potential for longer-distance dispersal events that fall outside of the study area. Our results may offer a glimpse into bumble bee dispersal including longer distances that would be difficult to measure directly.

Hippoglossus hippoglossus: From the RADseq data of Kess et al. (2021) we analyzed 11 individuals with 69,000 SNPs. Tagging studies find mean halibut movements greater than 100 km (Liu et al., 2019). However, the distance traveled by adults in search of food may be considerably larger than the quantity we wish to estimate which is proportional to the mean distance between birth location and parental birth location. Indeed, there is spatial structure distinguishing Atlantic halibut stocks due to spawning site fidelity (Shackell et al., 2021). Although in the case of halibut, the geographic area where the analyzed samples were collected may does not represent the spawning grounds, because of the long distances traveled by adults. Therefore, the observed sample locations—used to calculate the second input to `disperseNN`—are likely foraging locations that may differ significantly from the breeding locations. However, if assumptions about the size of the spawning area can be made, `disperseNN` provides a novel approach for inferring effective σ in foraging or migrating individuals for whom "home" locations are not known. Our estimate of 3.5 km (using the sampling width as the second input) could be close to the true dispersal distance if birth site fidelity is quite high. In another large marine species, *Diplodus sargus sargus*, natal dispersal distance was measured to be 11 km using otolith chemistry (Di Franco et al., 2012).

Crassostrea virginica: From the RADseq data of Bernatchez et al. (2019) we analyzed 13 individual eastern oysters with 7,097 SNPs. This species has larval dispersal (Vercaemer et al., 2010) and occasional adult translocations (Bernatchez et al., 2019). Our estimate of 1.2 km is much smaller than the previous estimate of 21.9 km (Rose et al., 2006). We offer several possible explanations for this discrepancy. We expect that oyster dispersal depends more on the strength and direction of local currents, rather than locomotion, and our training data did not include a mean "drift" component to dispersal. The previous estimate was from a different sample region, Chesapeake Bay, which likely has different local conditions than the coast of Canada where the samples that we analyzed were collected. Second, the previous estimate used microsatellite loci to estimate density in order to implement the Rousset method. Density is notoriously difficult to estimate from genetic data, so it would not be surprising if this step contributed to error. In contrast, `disperseNN` is designed to work around the unknown density parameter. However, we note that the marine species analyzed here potentially violate the two-dimensional habitat assumption of our model.

Canis lupus: From the RADseq dataset of Schweizer et al. (2016) we analyzed data from 13 individual wolves genotyped at 22,000 SNPs. Exceptionally good data exist on wolf dispersal from radio collars. A commonly reported value for this species is the distance traveled by adults that disperse between territories. For example, some estimates for this value include 98.1 km (Jimenez et al., 2017), 98.5 km (Kojola et al., 2006), and 147.0 km (Barry et al., 2020). However, not all individuals disperse from their natal territory. For example 50% and 47% of individuals dispersed between territories in Kojola et al. (2006) and Barry et al. (2020), respectively. Jimenez et al. (2017) reported more nuanced statistics: 18% of collared individuals had documented dispersal, survival was lower in dispersers, and not all dispersers reproduced. It is unclear how frequent breeding occurs *within* the natal pack; if 85-90% of reproduction occurred without movement between territories, then our estimate of 12.8 km might be reasonably close to the true, effective dispersal distance.

Helianthus: We analyzed two wild sunflower species from Todesco et al. (2020): *Helianthus petiolarus* ($n = 11$; 61,000 SNPs) and *H. argophyllus* ($n = 30$; 60,000 SNPs), with whole genome resequencing data. Wild sunflowers regularly outcross, therefore the estimated σ in part reflects pollinator distance, in addition to transport of seeds, e.g., by animals. Previously, Arias and Rieseberg (1994) reported the frequency of hybridization between cultivated and wild sunflowers at distances between 3 m and 1000 m; if we convert these hybridization-frequencies to counts of hybridization events, the mean distance of these pollination events was 156 m. The estimates from `disperseNN` were larger: 820m and 850m in *H. petiolaris* and *H. argophyllus*, respectively. These estimates may be reasonable if pollination occurs via bees, which can have foraging ranges greater than 1 km (Osborne et al., 2008; Visscher and Seeley, 1982). Studying foraging distance in pollinators is an active area of research, however Pasquet et al. (2008) used an exceptionally large study area and radio trackers to find a median flight distance of 720m in carpenter bees. Our estimates for the two analyzed *Helianthus* species were similar to each other.

Arabidopsis thaliana: From the whole genome resequencing dataset of the The 1001 Genomes Consortium (2016) we were able to analyze two sampling clusters from different geographic regions: Spain (142,000 SNPs, n=35) and Sweden (124,000 SNPs, n=84). Our σ estimates in these populations were 1,110 m and 360 m, which are considerably larger than the average distance that seeds fall from the parent plant; Wender et al. (2005) estimated that the average distance traveled by *A. thaliana* seeds with wind is less than 2 m. However, occasional long distance seed dispersal, e.g., via water or animals, and infrequent outcrossing via insect pollination may inflate the effective dispersal distance in this species. Outcrossing in *A. thaliana* has been estimated to be 3×10^{-3} (Abbott and Gomes, 1989). Importantly, *A. thaliana* is predominantly selfing and the analyzed samples are (naturally) inbred, which is misspecified by the current training set which did not include selfing. *A. thaliana* has experienced a known population expansion (Tyagi et al., 2016), and although we attempted to account for demographic history during training the true history of *A. thaliana* may not be well-represented by our simplistic range of population histories. There was a three-fold difference in estimated dispersal distance between the analyzed populations, perhaps due to local environmental differences between Spain and Sweden or different pollinator species.

Discussion

Dispersal estimation using deep learning

Understanding how organisms move across land or seascapes is critical for gaining a full picture of the forces shaping genetic variation (Wright, 1943; Kimura and Weiss, 1964; Barton et al., 2002). However, it remains difficult to confidently infer spatial population genetic parameters. Here we present a deep learning framework, `disperseNN`, for estimating the mean per-generation dispersal distance from population genetic data. There are several advantages of our method over existing population-genetics-based estimators, including improved accuracy for small to moderate sample sizes, accessible input data (unphased SNPs), and the ability to infer dispersal distance in the face of unknown model parameters, e.g., population density. These improvements open the door for using DNA to infer dispersal distance in non-model organisms where population density is unknown or identity-by-descent tracts are out of reach. Because `disperseNN` uses a form of simulation-based inference, analyses can be tailored for the particular study system, for instance detailed habitat maps and independent estimates for key model parameters can be readily incorporated.

Unlike previous genetics-based estimators that use geographic distances between individuals, our neural network does not see the relative spatial locations of individuals. This means that our neural network could in theory be applied to genetic data for which sampling locations are unavailable, or applied to adult individuals that have ranged far from their nesting or spawning area. However, to do so an estimate of the sampling width is required as input by `disperseNN`. Further, competing methods summarize the genotypes as genetic distances or identity-by-descent blocks between pairs of individuals. While these measurements are natural choices to focus on for analyzing dispersal, they inherently miss other information potentially available in the genotypes. The rate of dispersal affects not only pairwise genetic distances between individuals, but also population genetic variation more generally, such as nucleotide diversity, the site frequency spectrum, etc. (Battey et al., 2020b). `disperseNN`, by using a convolutional neural network with a complete genotype matrix as its input, is able to capture population genetic information from raw data as has been seen in a few prior contexts (e.g., Flagel et al., 2019; Sanchez et al., 2021; Gower et al., 2021). This in principle allows the network to see additional aspects of genetic variation—the distribution of allele frequencies, linkage disequilibrium if present, etc.—which has the potential to improve dispersal estimates. Indeed, `disperseNN` outperforms other, state-of-the-art tools, particularly when the sample size is small ($n = 10$).

Another strength of the deep learning approach is its versatility. In particular, `disperseNN` can be used with unphased SNPs and small sample sizes, which makes it applicable for a variety of genomic dataset types. In contrast, recently developed tools for dispersal estimation require identity-by-descent blocks as input (Ringbauer et al., 2017; Al-Asadi et al., 2019). Although these methods perform well when high quality data is available, phasing and identity-by-descent inference in non-human genomes is a considerable challenge, especially for RADseq. Unphased SNPs, on the other hand, are more widely available. Our approach addresses this gap in available methods by facilitating unphased data.

Next, our inference framework allows dispersal inference without *a priori* knowledge of important nuisance parameters, namely population density and the habitat size. In contrast, the commonly used Rousset method

requires an independent estimate for population density in order to infer dispersal distance. Our supervised learning approach can learn to predict σ in the face of unknown density, which is achieved by exposing the network to training datasets with various densities. Through this procedure, `disperseNN` successfully learned to estimate σ in test datasets regardless of density, conditioned on true density being within the training distribution. While that is so, we still observed misspecification for large, out-of-distribution densities, which caused the network to overestimate σ. The same approach can be used if other parameters are uncertain, for example the size of the habitat. On the other hand, if independent estimates for some parameters or better-informed "priors" are available, then training can be customized to reflect the known parameters.

Thus far we have focused on indirect estimation of dispersal distance, without measurements of how far individuals move. For a review of other genetic techniques for estimating dispersal distance, including direct and indirect methods, see Broquet and Petit (2009). Recently, two studies have used close-kin mark-recapture approaches for estimating dispersal distance, which were applied to mosquito species (Jasper et al., 2019; Filipović et al., 2020). Close kin mark recapture uses the genome of a close relative to represent a "recapture", thereby skipping the need to physically recapture individuals. These promising new methods estimate dispersal distance by modeling the spatial distribution of close kin. In theory, our approach may offer advantages over close kin mark recapture: `disperseNN` aims to estimate *effective* dispersal, has no requirement for close kin to be captured together, and works with small sample sizes ($n = 10$). The ability to capture kin relies on a sample size that is a sufficiently large proportion of the local population size, which is not always feasible.

Limitations

Although training on simulated data allows great flexibility, the simulation step was also a limitation for the current study. In particular, generating the training data for our pre-trained network involved very long computational run times and large memory requirements: up to 175 gigabytes of RAM and two weeks of run time for the largest parameterizations of individual simulations. Shortcuts were used to reduce simulation time, including: only 100 generations of spatial simulation, and sampling multiple times from each simulated population (see Materials and Methods). Of course, if new training data are generated for a population that is comparatively small, then the simulation burden will be smaller.

As with many statistical approaches, `disperseNN` has limited ability to generalize outside of the range of parameter values on which it was trained. Although we successfully dealt with individual nuisance parameters, for example by exposing the model to training datasets with varying density, it was unable to extrapolate to out-of-sample data. If the test data had very large population density—higher than the network had seen during training—σ was overestimated. Likewise, prediction error increased if the test data had a larger spatial sampling area than the network saw during training. Therefore we expect the pre-trained model from our empirical analysis to be most accurate for smaller spatial samples from smaller populations—parameters that fall inside the training range—while applications to larger populations may be more questionable. In fact, it is generally recommended to restrict the sampling area to a small region when estimating σ to avoid issues with environmental heterogeneity and patchy habitats (Broquet and Petit, 2009; Shipham et al., 2013). However, a sufficiently large sampling area is required to infer large σ.

Another potential issue with our approach is complex demographic history. As demographic perturbations leave a footprint in contemporary genetic variation, demography may bias estimates of σ for a neural network trained with a particular history, e.g., constant N. This issue is by no means unique to our analysis. Leblois et al. (2004) showed that estimates using Rousset's technique mostly reflected past demographic values rather than recent population density. We attempted to address this in our analysis, by simulating under random two-epoch models. This approach was accurate for test data with a similar two-epoch history. However it also suggests that different, more complex demography may reduce accuracy, for example a more extreme bottleneck than was simulated in training, fluctuating N, pulse admixture, or perhaps population structure not captured in our simulations (e.g., barriers to dispersal or range expansion). Identity-by-descent based methods may alleviate the effect of ancestral population structure because long identity-by-descent tracts originate from the recent past (Barton et al., 2013). Similar to demographic history, other model misspecifications such as complex habitats and environmental heterogeneity could also be sources of error for estimation using our method.

Likewise, in our model dispersal is uniform across space. This assumption may be nearly true—or, at

least useful—for certain applications, particularly if the sampling area is small. However, in reality we expect dispersal to vary across space due to ecology: for example, mountain ranges will prohibit dispersal for many species. Alternatively, suitable habitat is often discontinuous, and dispersal between patches may be different than within patches. Likewise, heterogeneous habitat can generate source-sink dynamics between patches. Existing methods that infer heterogeneous dispersal surfaces across space (Petkova et al., 2016; Al-Asadi et al., 2019) have limitations including (i) estimating relative differences in dispersal as opposed to the magnitude of dispersal, or (ii) requiring identity-by-descent data as input.

When we included multiple nuisance parameters (Figure 4; Parameter Set 11), the MRAE was larger than that of experiments with only one or zero nuisance parameters (Figure 3; e.g., Parameter Set 3). This difference can be partly explained by the larger number of parameters with potential to confound. In addition, the *range* of values explored for σ, as well as for nuisance parameters, were orders of magnitude larger than those of the other experiments.

Interpretation of empirical findings

We estimated σ in a diverse set of organisms using publicly available datasets. These included both whole genome shotgun and RADseq—i.e., variations on standard RADseq (Baird et al., 2008) or genotyping-by-sequencing protocols (Elshire et al., 2011). Rather than simulate scenarios that would be appropriate to each species independently, we trained a single `disperseNN` model designed to estimate σ without *a priori* knowledge of density, ancestral population size, or species range.

The majority of empirical results from `disperseNN` were sensible, however our estimates for *A. thaliana*—particularly in the population located in Spain—are likely overestimates, in part due to the lack of selfing in our training simulations. *A. thaliana* had levels of heterozygosity and inbreeding that were outside the range of values observed in the training set, a feature reflected in the Mahalanobis distances between training and prediction sets. In the future, `disperseNN` might be better tuned to analyze selfing species, but this would require simulating additional training data and subsequent validation steps.

Our approach led to consistently larger dispersal estimates than mark-recapture experiments. Mark-recapture data was available for three of the analyzed taxa—white footed mouse, *Bombus*, and *Anopheles*. However the mark-recapture estimates for *Anopheles* are not ideal, as they represent only adult-travel distances, i.e., foraging distance. In contrast, the measurements from bumble bees (Carvell et al., 2017; Lepais et al., 2010) and mice (Keane, 1990; Jacquot and Vessey, 1995) are particularly relevant, as they measure the distance traveled by queen bees from the original hive or individual mice between birth location and adult territory. In all three cases our estimate was larger than the mark-recapture calculation, which suggests either an upward bias in the `disperseNN` output or underestimation in the mark-recapture estimates. In each mark-recapture study the geographic recapture area was smaller than the sampling area we provided to `disperseNN`. It is likely that long-distance dispersers, even if less common, are missed during the recapture step, which would bias the inferred dispersal distance downward in direct, mark-recapture studies.

Population genetics for spatial ecology

An understanding of dispersal is critical for preserving biodiversity (Driscoll et al., 2014). Dispersal is one of the main factors controlling metapopulation dynamics (Leibold et al., 2004), as well as the total population size and whether a population persists (Gadgil, 1971). Therefore, dispersal estimates are critical for choosing appropriate settings in population viability analyses (Akçakaya and Brook, 2008). Likewise, geographic habitat shifts are ongoing for many species, and species' survival may thus depend on their ability to disperse fast enough to follow rapidly changing local conditions (Wiens, 2016). Thus, obtaining values for dispersal distance are important for species distribution modeling which is used to project future species ranges (Wiens et al., 2009). In the comprehensive review of Driscoll et al. (2014), the authors present a list of 28 applications for which dispersal values were needed in conservation management, and report several independent calls for improved dispersal information and dispersal inference methods (Broquet and Petit, 2009; Hadley and Betts, 2012; Ceballos et al., 2009; Kingsford et al., 2009; Sutherland et al., 2006; Noss et al., 2009; Pullin et al., 2009).

Characterizing dispersal is also important for managing animal populations relevant to human health. For example, in the fight against malaria we must identify migration corridors and source-sink dynamics

in mosquito vector species to allocate pesticide treatment and to predict the spread of genetic variants conveying insecticide resistance (Clarkson et al., 2020). Understanding dispersal is particularly important for modeling and implementing gene-drive strategies (Champer et al., 2021; Beaghton and Burt, 2022; North et al., 2013, 2019, 2020; Beaghton et al., 2016, 2017) for controlling the spread of mosquito-borne diseases including malaria.

Direct methods such as radio tracking or genetic identification may provide near-perfect measurements of dispersal within the generation or generations analyzed. However it is often more valuable to know the expected dispersal distance over many generations, conditional on survival and successful reproduction of the dispersing individuals. For example, the day-to-day foraging distance or seasonal migration distances traveled by adults may differ from the effective dispersal distance. Direct methods such as mark-recapture are often expensive and as a result are limited to relatively small geographic areas, which may ignore long distance movement and bias the resulting estimate. Population genetic tools therefore complement direct methods for improving our understanding of dispersal.

Materials and Methods

Simulations

Training datasets were simulated using an individual-based, continuous-space model based on that of Battey et al. (2020b). The simulation is initialized with hermaphroditic, diploid individuals distributed randomly on a square habitat. The life cycle of an individual consists of stages for dispersal, reproduction, and mortality. Each offspring disperses from the maternal parent's location by an independent random displacement in each dimension that is Gaussian distributed with mean zero and standard deviation σ. The mate of each individual in each time step is selected randomly, with probability proportional to the Gaussian density with mean zero and standard deviation σ, up to a maximum of 3σ units in space. The number of offspring per mating is Poisson distributed with mean $\frac{1}{4}$. Competitive interactions with neighboring individuals affect the probability of survival, allowing the total population size to fluctuate around an equilibrium. Specifically, individuals at distance d compete with strength $g(d)$, leading to a cumulative interaction strength for individual i of $n_i = \sum_j g(d_{ij})$, where d_{ij} is the distance between individuals i and j. These competitive interactions extend to a distance of 3σ. The probability of survival for individual i, is $p_i = min\left(0.95, \frac{1}{1+n_i/(K(1+L))}\right)$, where K and L are parameters that are approximately equal to the the carrying capacity per unit area and the average lifetime at equilibrium, respectively. Thus, a single parameter, σ, is used to control three different processes simultaneously: dispersal, mating, and competition. Edge effects are avoided by decreasing individual fitness proportional to the square root of distance from the habitat edges in units of σ. Offspring whose proposed location falls outside of the bounds of the habitat are not generated. This model was implemented in SLiM 3.7 (Haller and Messer, 2019). We used a genome length of 10^8 bp and recombination rate 10^{-8} crossovers per bp.

After the completion of the spatial, forward-in-time SLiM simulation, initial genetic diversity was produced using a coalescent simulation in msprime, known as "recapitation" (Kelleher and Lohse, 2020). This strategy was necessary to reduce computation time to manageable levels, as the coalescent stage of the simulation is much faster than the spatially-explicit portion. The ancestral N_e was set to the "present day" census population size for recapitation. This portion of the simulation proceeded until all genealogical trees had coalesced. Thus, the complete simulation involves random mating for older generations equivalent to a Wright Fisher model, with a number of recent generations that are spatially explicit (Table 2). Most of our experiments used 100,000 spatial-SLiM generations. However, to facilitate larger simulations for the multiple nuisance parameters experiment (Parameter Set 11), we ran only 100 generations of spatial SLiM due to computational limitations. We found that `disperseNN` can predict σ from full-spatial test data after training on simulations with only 100 spatial generations, although σ was moderately underestimated when testing with larger numbers of spatial generations (Figure S9). To simulate population size changes, we recapitated with msprime as before, but included an instantaneous decline or expansion between 100 and 100,000 generations in the past.

Other model parameters varied between experiments and the relevant parameter ranges are described

in Table 3. Population density is one quantity that is focused on in our study, however density is an emergent property of our simulation rather than a model setting. To control population density we vary the carrying capacity per unit area, K, in the simulation which is the main determinant of density. In practice, mean density fluctuates moderately. When the specified size of the spatial sampling window was smaller than the full habitat, the position of the sampling window was chosen randomly, with x and y each distributed uniformly (Figure 5), excluding edges. The amount of edge cropped was either set to (i) σ for each simulation, or (ii) the maximum of the simulated σ range for the whole training set, depending on which simulation parameters were free to vary; the latter was necessary to avoid information leakage during training. Individuals were sampled randomly from within the sampling window.

Params.	Description	Sims.	Training	Spatial gen.	n	SNPs	Phased
1	Comparing estimators	1,000	50,000	100,000	10 and 100	$2.5 \times 10^5, 5 \times 10^5$	Y
2	Baseline	1,000	50,000	100,000	100	5,000	Y
3	Variable density	1,000	50,000	100,000	100	5,000	Y
4	Large density	1,000	50,000	100,000	100	5,000	Y
5	Demographic history	1,000	50,000	1,000	100	5,000	Y
6	Extreme ΔN change	1,000	50,000	1,000	100	5,000	Y
7	Variable habitat size	1,000	50,000	100,000	50	5,000	Y
8	Large habitat size	1,000	50,000	100,000	50	5,000	Y
9	Variable sampling width	1,000	50,000	100,000	100	5,000	Y
10	Large sampling width	1,000	50,000	100,000	100	5,000	Y
11	Multiple nuisance par.	2,300	100,000	100	U-int(10,100)	5,000	N

Table 2: Analysis parameters. The "Params." column lists the identifier for the parameter set, which is referenced in the main text. "Description" is a brief description of the parameter set. "Sims." is the number of true replicates, i.e., SLiM simulations, represented in training. "Training" is the size of the total training set after drawing multiple samples from each simulation. "Spatial gen." is the number of spatial generations simulated in SLiM. "n" is the sample size. "SNPs" is the number of SNPs used in training. "Phased" describes whether the data were phased or not for training.

Params.	Description	σ	K	ΔN	Habitat width	Samp. width	Edge
1	Comparing estimators	U(0.2, 3.0)	5	constant	50	1	3
2	Baseline	U(0.2, 3.0)	5	constant	50	1	3
3	Variable density	U(0.2, 3.0)	log-U(0.1, 20.0)	constant	50	1	3
4	Large density	U(0.2, 3.0)	U(20.0, 40.0)	constant	50	1	3
5	Demographic history	U(0.2, 3.0)	5	$\begin{cases} U(\frac{1}{5}, 1) \\ U(1, 5) \end{cases}$	50	1	3
6	Extreme ΔN change	U(0.2, 3.0)	5	$\begin{cases} U(\frac{1}{10}, \frac{1}{5}) \\ U(5, 10) \end{cases}$	50	1	3
7	Variable habitat size	U(0.2, 3.0)	2	constant	U(15, 150)	1	σ
8	Large habitat size	U(0.2, 3.0)	2	constant	U(150, 300)	1	σ
9	Variable sampling width	U(0.2, 3.0)	5	constant	50	U(0.2, 0.8)	σ
10	Large sampling width	U(0.2, 3.0)	5	constant	50	U(0.8, 1.0)	σ
11	Multiple nuisance par.	log-U(10^{-3}, 10^2)	log-U(10^{-3}, 10^4)	$\begin{cases} U(\frac{1}{5}, 1) \\ U(1, 5) \end{cases}$	log-U(2, 10^3)	U(0.0, 1.0)	σ

Table 3: Parameter distributions used for simulation. The "Params." column lists the identifier for the parameter set, which is referenced in the main text. "Description" is a brief description of the parameter set. "σ" is the distribution of the dispersal parameter. "K" is the major determinant of population density. "ΔN" describes the history of population size change: for rows with braces, a random multiplier was chosen from one of two uniform distributions, each with probability 0.5. The ancestral N_e was set to the multiplier $\times$ present day N. "Habitat width" is for the full habitat. "Samp. width" is the width of the sampling area as a proportion of the full habitat width. "Edge" is a distance from each side of the habitat that was excluded from sampling to avoid edge effects .

To obtain genetic data, neutral mutations were superimposed on the tree sequences using msprime v1.0

(Baumdicker et al., 2022) until a predetermined number of SNPs, m, were obtained (Table 2). Specifically, we started by simulating mutations with a very small mutation rate, 10^{-15}. Next, we increased the mutation rate by 10x, and threw on additional mutations with the updated mutation rate. The latter two steps were iterated several times until at least m mutations had been obtained. When at least m SNPs had been added, m SNPs were sampled to represent the genotype matrix input to `disperseNN`. The result of this procedure is that the genotype matrix for each simulated dataset contains the same number of SNPs, m, regardless of the actual number of variable sites in the sampled individuals, and irrespective of mutation rate, and thanks to the Poisson nature of neutral mutations is equivalent to having simulated with a higher mutation rate and randomly selected m variable sites. For some analyses, multiple samples were drawn from the same simulated tree sequence to save computation time; these cases are noted in Table 2. This strategy allows for large training sets to be generated from a smaller number of starting simulations.

The input for `disperseNN` consists of two things: the width of the spatial sampling area, and a genotype matrix, having one row for each SNP and one or two columns per individual depending on the phasing designation. If phased, the genotype matrix contained two columns per individual, randomly ordered, with 0s and 1s encoding minor and major alleles, respectively. If unphased, the genotype matrix contained one column per individual with genotypes encoded as 0s, 1s, and 2s, representing the count of the minor allele. In order to facilitate various sample sizes in real applications, our pre-trained model used a random sample size during training, $10 \leq n \leq 100$, with zero padding out to 100 columns. To obtain the second input, we used the furthest distance between pairs of samples as the sampling width. The training targets are the true σ's, log-transformed. Thus, the output from the CNN is in log space (`disperseNN` exponentiates the result before writing the predictions).

In generating training data for the pre-trained network, we sought to explore a large parameter range: each parameter varied over several orders of magnitude (Parameter Set 11). However, swaths of parameter space described by the ranges in Table 3 were not represented in the training data, due to the following logistical hurdles. First, simulations where the population died were not included in the training set. The excluded simulations had small carrying capacity and small habitat size, or small habitat size and large σ, for example. Next, some simulations could not be run due to computational constraints: maximum RAM of 175 gigabytes and two-week wall time on our computing cluster. For example, combinations of large carrying capacity and large habitat size were not simulated. As a result, only 12% of attempted simulations were included in training, and for each parameter the *realized* distribution—representing successful simulations—differed from the distribution from which the model settings were drawn (Figure S6), which had been uniform in log space.

CNN architecture and training

Tensorflow (Abadi et al., 2016) and Keras (`https://github.com/keras-team/keras`) libraries were used to develop `disperseNN`. The first input tensor, the genotype matrix, goes through successive convolution and pooling layers, a strategy that is characteristic of CNNs (Figure 1). We adjusted the number of convolution and pooling layers based on the size of the genotype matrix: the number of convolution layers assigned was equal to $\text{floor}(\log_{10}(\text{number of SNPs}))-1$. The filter size of successive convolution layers was 64 for the first layer, and 44 larger for each successive layer. The convolution layers are one-dimensional, such that the convolution kernel spans all individuals (columns) and two SNPs (rows), with stride size equal to one. Average pooling layers were also one dimensional, spanning all individuals and 10 SNPs. After the convolutional portion of the network, the intermediate tensor was flattened and put through three fully connected layers each with 128 units and rectified linear unit (ReLU) activation. A second input branch was used for the sampling area. This input tensor with size = 1 was concatenated with the preceding branch, then subjected to a 128-unit dense layer with ReLu. Finally, a dense layer with linear activation was applied which outputs a single value, the estimate for σ.

During training we held out 20% of the training set for computing a validation-loss between epochs. We used a batch size of 40, mean squared error loss, and the Adam optimizer. The learning rate was initialized as 10^{-3}. The "patience" hyperparameter determines both the length of training, and learning rate adaptations during training: after a number of epochs equal to patience/10 without improvement in validation loss the learning rate is halved, and training proceeds until a number of epochs equal to patience pass without improvement in validation loss. Patience was set to 100 for all training runs excluding the pre-trained model.

For the pre-trained model, we explored a grid of different hyperparameter settings: patience values of 10, 20, 30, 40, and 50; initial learning rates of 10^{-4}, 10^{-3}, and 10^{-2}; and dropout proportions of 0, 0.1, 0.2, and 0.3. We landed on settings that consistently gave the lowest MRAE: patience = 10, initial learning rate of 10^{-3}, and 0 dropout.

Comparison with other estimators

The Rousset method uses the observation that under certain assumptions, then $b = 1/4\pi D\sigma^2$ (and recall D is the effective density). b is the slope of the least squares linear fit of $a_r/(1-a_r)$ to geographic distance, where a_r is a measure of genetic differentiation between two individuals analogous to F_{st}, where from Rousset (2000), a_r for a pair of individuals, $\mathscr{P}$, can be estimated as $\hat{a}^* = \frac{(2SS_{b(\mathscr{P})} - SS_{W(\mathscr{P})})P}{2\sum_{k=1}^{P} SS_{W(k)}}$, where $SS_{b(\mathscr{P})}$ is the sum of squared differences between the two individuals' genotypes, $SS_{W(\mathscr{P})}$ is the sum of squared differences between genomes within the individuals, P is the total number of pairs of individuals in the sample, and $\sum_{k}^{P} SS_{W(k)}$ are within individual differences summed over the P different pairs of individuals. We applied Rousset's method to the same genotypes and sample locations as for `disperseNN`. The values for D used with this method were calculated after excluding the edges of the habitat which have reduced density: the census size, N, was counted after excluding individuals within an edge width, E, from any side of the habitat of width W, thus $D = N/(W-2E)^2$.

A second comparison was made with `IBD-Analysis` (Ringbauer et al., 2017). The authors used the distribution of identity-by-descent tract lengths shared between individuals to estimate σ. They derived analytical formulas describing how isolation-by-distance shapes identity-by-descent tracts and provided an inference scheme that uses maximum likelihood to fit these formulas. For our comparison, we extracted perfect identity-by-descent tracts directly from the tree sequences output from SLiM. Specifically, for each pair of individuals, for each combination of chromosomes between the individuals, we simplified the tree sequence to represent only the recombination history between the two chromosomes, and extracted segments that were inherited from a common ancestor without recombination. These were the identity-by-descent tracts used as input for the Analyze-IBD program, which was obtained from `https://git.ist.ac.at/harald.ringbauer/IBD-Analysis`. Separately, we inferred identity-by-descent tracts in the simulated data using an empirical tool, `Refined IBD` (Browning and Browning, 2013), and used the inferred identity-by-descent as input `IBD-Analysis`. For the latter analysis a mutation rate of 10^{-8} was applied and all variant sites were included in the identity-by-descent inference step, with other parameters the same as in Parameter Set 1.

Empirical data

To demonstrate the utility of `disperseNN`, we applied it to preexisting publicly available empirical datasets that have the following criteria: spatially distributed genetic data, latitude and longitude metadata available, ten or more sampling locations, sampling area less than 1000 km, at least 5,000 biallelic SNPs, and a ready-to-plug-in SNP table that had been processed and filtered by the original authors. For some datasets with overall sampling width more than 1000 km, we were able to subset for a smaller cluster of sample locations (see details specific to each dataset below). When multiple individuals were sampled from the same location we chose one random individual from each location, in order to better match the sampling scheme used in generating training data. SNP tables were converted to genotype matrices after minimal processing: we removed indels and sites with only one, or more than two, alleles represented in the sampled subset. We required all sampled individuals to be genotyped to retain a SNP, except when we note otherwise—see details specific to each dataset below.

Mosquito data were downloaded following instructions from `https://malariagen.github.io/vector-data/ag3/download.html`. We used a dense cluster of sampling localities in Cameroon that had been identified as *Anopheles gambiae*. Individual VCFs were merged using `bcftools` (v1.14). Chromosomes 3L and 3R were analyzed; 2L and 2R were excluded due to previously reported large inversions (Lobo et al., 2010; Riehle et al., 2017).

Arabidopsis data was downloaded from `https://1001genomes.org/data/GMI-MPI/releases/v3.1/` as a single VCF. Two conspicuous geographic clusters were chosen from Sweden and Spain to minimize the geographic sampling area. All five chromosomes were analyzed.

Sunflower data was downloaded from cloud storage following instructions from `https://rieseberglab.github.io/ubc-sunflower-genome/documentation/`. Geographic clusters of sampling localities were identified in Texas (*Helianthus argophyllus*) and on the border of Kansas and Oklahoma (*H. petiolaris*). Individual VCFs were merged into multi-sample VCFs for each of the two species. Chromosomes 1-17 were analyzed, excluding a number of unplaced scaffolds.

VCFs for oyster (*Crassostrea virginica*; Bernatchez et al. (2019)), bumble bee (*Bombus*; Jackson et al. (2018)), Atlantic halibut (*Hippoglossus hippoglossus*; Kess et al. (2021)), white-footed mouse (*Peromyscus leucopus*; Munshi-South et al. (2016)), Réunion grey white-eye (*Zosterops borbonicus*) and Réunion olive white-eye (*Zosterops olivaceus*; Gabrielli et al. (2020)), and wolf (*Canis lupus*; Schweizer et al. (2016)) were downloaded directly from The Dryad Digital Repository. Clusters of sample locations were chosen in each dataset to maximize sampling density. In the datasets from *Bombus vosnesenskii*, *Peromyscus leucopus*, *Zosterops borbonicus*, and *Zosterops olivaceus*, we allowed as few as 85%, 60%, 90%, and 90% of individuals to be genotyped to retain a SNP, respectively, and missing genotypes were filled in with the major variant.

To calculate the width of the sampling window for empirical data, we calculated the geodesic distance between each pair of individuals using the package `geopy` with the WGS84 ellipsoid. This distance represents the shortest path on the surface of the Earth between points. The longest distance between pairs of sample locations was used as the sampling width, which we provided in kilometers to `disperseNN`.

Acknowledgements

We are thankful to Harold Ringbauer for assistance with the methods comparison analysis. This work was funded by NIH awards F32GM146484 to CCS and R01HG010774 to ADK. Computation was done using the University of Oregon's cluster, Talapas, with help from the Research Advanced Computing Services team. We thank members of the Kern-Ralph Co-lab for valuable input along the way.

Data availability

The `disperseNN` code is available on GitHub at the following link: `https://github.com/kr-colab/disperseNN`.

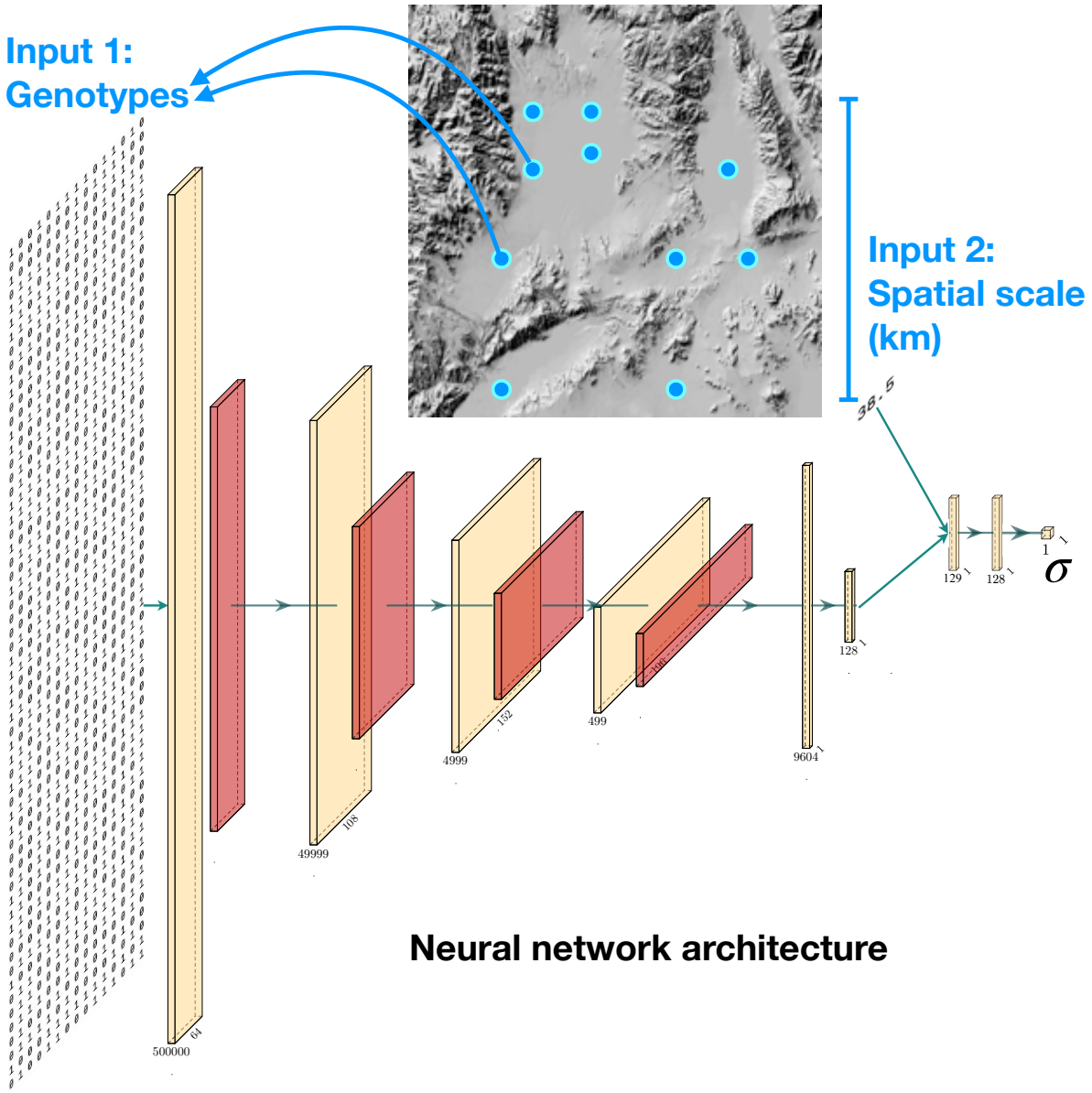

Figure 1: Diagram of the analysis workflow. Blue points are hypothetical sample locations on a geographic map. Rectangular neural network layers are 1D-convolution and average-pooling layers; columnar layers are fully connected layers. The two input branches are concatenated into a single, intermediate tensor. Neural network schematic generated using PlotNeuralNet (https://github.com/HarisIqbal88/PlotNeuralNet).

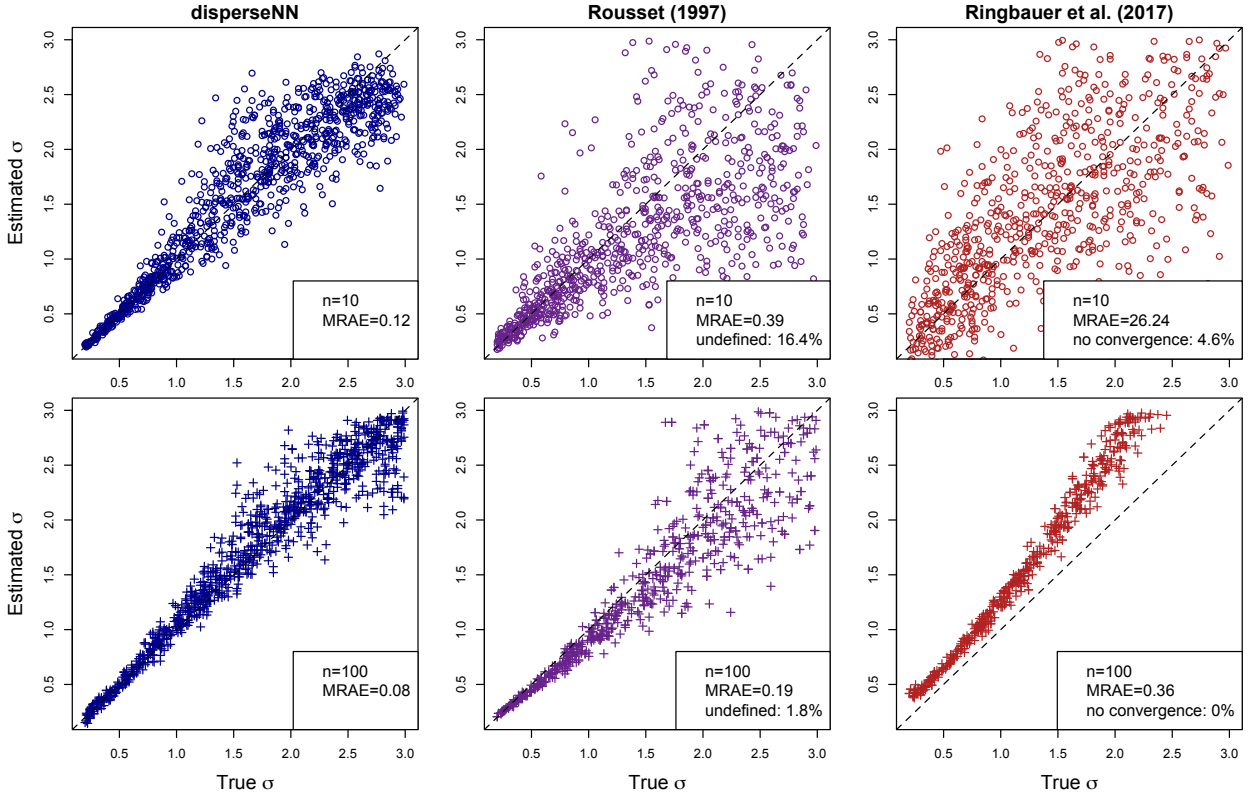

Figure 2: Comparison with existing methods (Parameter Set 1). Here `disperseNN` is compared with the Rousset method and `IBD-Analysis`, using two different numbers of sampled genomes, $n = 10$ (top row) and $n = 100$ (bottom row). Estimates greater than 3 are excluded from plots but are included in the MRAE calculation. Moreover, the Rousset method produced undefined output for 16.4% and 1.8% of $n = 10$ and $n = 100$ datasets, respectively; these data are not reflected in the MRAE calculation. Likewise `IBD-Analysis` did not converge for 4.6% of the $n = 10$ datasets. The dashed lines are $y = x$.

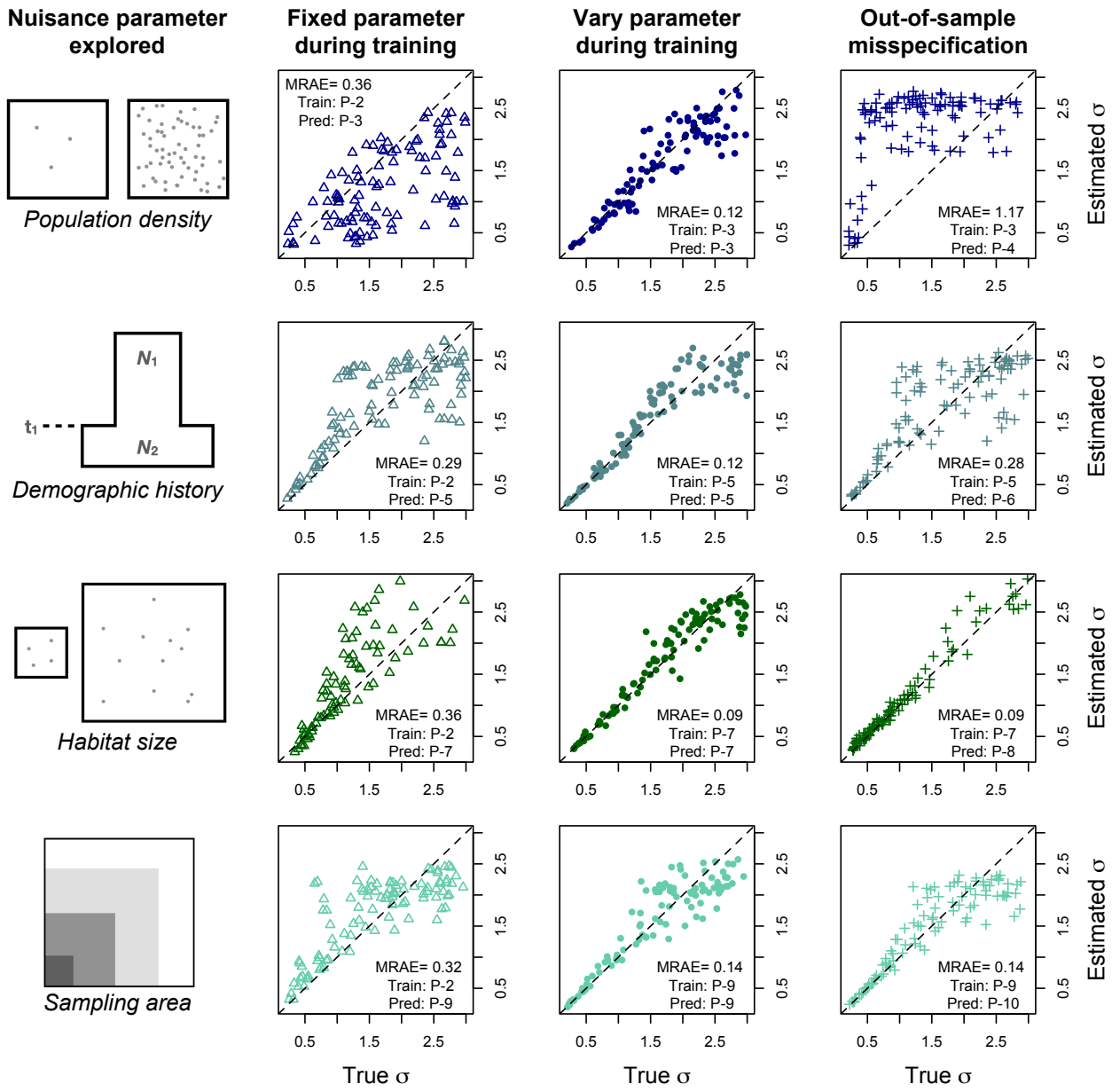

Figure 3: Column 1. Cartoons of unknown parameters that may lead to model misspecification. Column 2. The unknown parameter was fixed during training, but testing was performed on data with different values of the parameter. Column 3. The unknown parameter was *varied* during training, and testing was performed on data from the same distribution. Column 4. The unknown parameter was varied during training, but testing was performed on out-of-sample values, i.e., larger values than were seen during training. The dashed lines are $y = x$. Outliers greater than 3 are excluded from the fixed-habitat-size plot. "Train: P" and "Pred: P" refer to the Parameter Sets used for training and testing, respectively. The third row, third column plot has lower MRAE than the baseline model due to using a smaller carrying capacity, which was chosen to alleviate computation time.

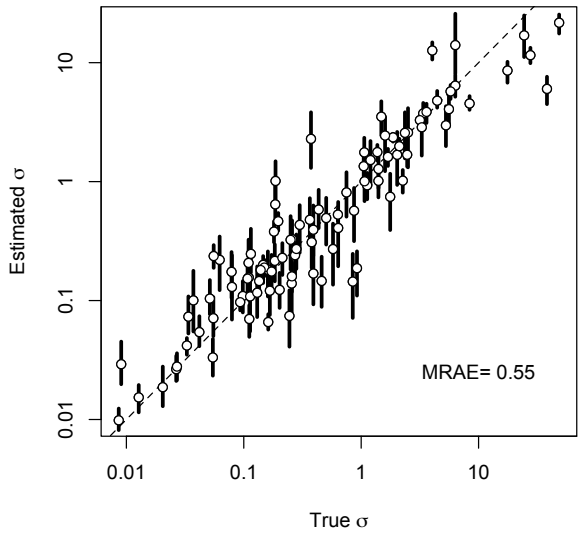

Figure 4: Validation of the pre-trained model (Parameter Set 11). Shown are 100 test datasets, each generated from an independent simulation. Open points indicate the mean estimate from 1000 subsamples of 5,000 SNPs drawn from each dataset. Also depicted is the range of estimates from the middle 95% of subsamples. The dashed line is $y = x$. Note the log scale.

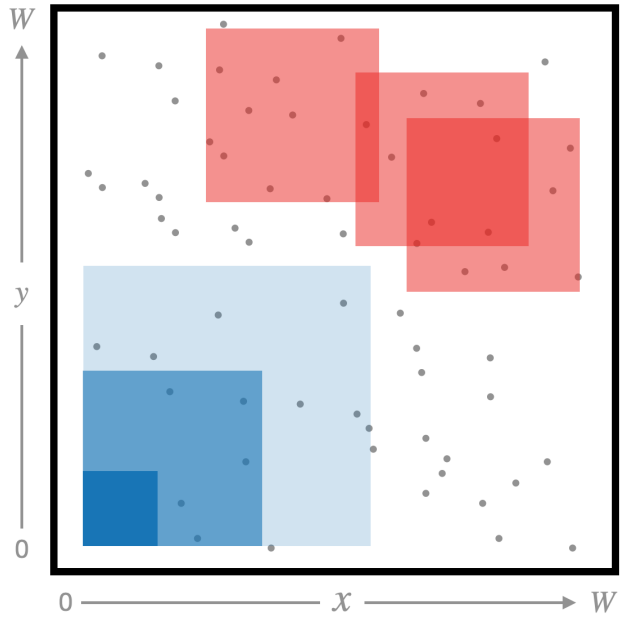

Figure 5: Cartoon showing different sampling strategies. The black box represents the full simulated habitat. For some experiments, we both (i) varied the width of the square sampling window—blue boxes show examples of differing sampling widths—, and (ii) assigned a uniform-random position for the sampling window—red boxes show different positions for the sampling window.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Richard J Abbott and Mioco F Gomes. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. *Heredity*, 62(3):411–418, 1989.

H Resit Akçakaya and Barry W Brook. Methods for determining viability of wildlife populations in large landscapes. *Models for planning wildlife conservation in large landscapes*, pages 449–472, 2008.

Hussein Al-Asadi, Desislava Petkova, Matthew Stephens, and John Novembre. Estimating recent migration and population-size surfaces. *PLoS genetics*, 15(1):e1007908, 2019.

Dulce M Arias and Loren H Rieseberg. Gene flow between cultivated and wild sunflowers. *Theoretical and Applied genetics*, 89(6):655–660, 1994.

Soheil Baharian, Maxime Barakatt, Christopher R Gignoux, Suyash Shringarpure, Jacob Errington, William J Blot, Carlos D Bustamante, Eimear E Kenny, Scott M Williams, Melinda C Aldrich, et al. The great migration and African-American genomic diversity. *PLoS genetics*, 12(5):e1006059, 2016.

Nathan A Baird, Paul D Etter, Tressa S Atwood, Mark C Currey, Anthony L Shiver, Zachary A Lewis, Eric U Selker, William A Cresko, and Eric A Johnson. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, 3(10):e3376, 2008.

Timothy Barry, Eliezer Gurarie, Farid Cheraghi, Ilpo Kojola, and William F Fagan. Does dispersal make the heart grow bolder? Avoidance of anthropogenic habitat elements across wolf life history. *Animal Behaviour*, 166:219–231, 2020.

Nicholas H Barton. The dynamics of hybrid zones. *Heredity*, 43(3):341–359, 1979.

Nick H Barton, Frantz Depaulis, and Alison M Etheridge. Neutral evolution in spatially continuous populations. *Theoretical population biology*, 61(1):31–48, 2002.

Nick H Barton, Alison M Etheridge, Jerome Kelleher, and Amandine Véber. Inference in two dimensions: allele frequencies versus lengths of shared sequence blocks. *Theoretical population biology*, 87:105–119, 2013.

Christopher J Battey, Peter L Ralph, and Andrew D Kern. Predicting geographic location from genetic variation with deep neural networks. *ELife*, 9:e54507, 2020a.

Christopher J Battey, Peter L Ralph, and Andrew D Kern. Space is the place: effects of continuous spatial structure on analysis of population genetic data. *Genetics*, 215(1):193–214, 2020b.

Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229, 2022.

Andrea Beaghton, Pantelis John Beaghton, and Austin Burt. Gene drive through a landscape: reaction–diffusion models of population suppression and elimination by a sex ratio distorter. *Theoretical population biology*, 108:51–69, 2016.

Andrea Beaghton, Andrew Hammond, Tony Nolan, Andrea Crisanti, H Charles J Godfray, and Austin Burt. Requirements for driving antipathogen effector genes into populations of disease vectors by homing. *Genetics*, 205(4):1587–1596, 2017.

PJ Beaghton and Austin Burt. Gene drives and population persistence vs elimination: the impact of spatial structure and inbreeding at low density. *Theoretical Population Biology*, 2022.

Simon Bernatchez, Amanda Xuereb, Martin Laporte, Laura Benestan, Royce Steeves, Mark Laflamme, Louis Bernatchez, and Martin A Mallet. Seascape genomics of eastern oyster (*Crassostrea virginica*) along the Atlantic coast of Canada. *Evolutionary Applications*, 12(3):587–609, 2019.

Gideon S Bradburd and Peter L Ralph. Spatial population genetics: it's about time. *Annual Review of Ecology, Evolution, and Systematics*, 50:427–449, 2019.

Thomas Broquet and Eric J Petit. Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology, Evolution, and Systematics*, 40:193–216, 2009.

Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

Claire Carvell, Andrew FG Bourke, Stephanie Dreier, Stephen N Freeman, Sarah Hulmes, William C Jordan, John W Redhead, Seirian Sumner, Jinliang Wang, and Matthew S Heard. Bumblebee family lineage survival is enhanced in high-quality landscapes. *Nature*, 543(7646):547–549, 2017.

Gerardo Ceballos, Mariana M Vale, Cristian Bonacic, Julio Calvo-Alvarado, Rurik List, Nora Bynum, Rodrigo A Medellín, Javier A Simonetti, and Jon Paul Rodríguez. Conservation challenges for the Austral and Neotropical America section. *Conservation biology*, 23(4):811–817, 2009.

Jackson Champer, Isabel K Kim, Samuel E Champer, Andrew G Clark, and Philipp W Messer. Suppression gene drive in continuous space can result in unstable persistence of both drive and wild-type alleles. *Molecular Ecology*, 30(4):1086–1101, 2021.

Chris S Clarkson, Alistair Miles, Nicholas J Harding, Eric R Lucas, Christopher J Battey, Jorge Edouardo Amaya-Romero, Andrew D Kern, Michael C Fontaine, Martin J Donnelly, Mara KN Lawniczak, et al. Genome variation and population structure among 1142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*. *Genome research*, 30(10):1533–1546, 2020.

The 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2):481–491, 2016.

The Anopheles Gambiae 1000 Genomes Consortium. Ag1000G phase 3 SNP data release, 2021. https://www.malariagen.net/data/ag1000g-phase3-snp.

Antonio Di Franco, Bronwyn M Gillanders, Giuseppe De Benedetto, Antonio Pennetta, Giulio A De Leo, and Paolo Guidetti. Dispersal patterns of coastal fish: implications for designing networks of marine protected areas. *PLoS One*, 7(2):e31681, 2012.

Don A Driscoll, Sam C Banks, Philip S Barton, Karen Ikin, Pia Lentini, David B Lindenmayer, Annabel L Smith, Laurence E Berry, Emma L Burns, Amanda Edworthy, et al. The trajectory of dispersal research in conservation biology. Systematic review. *PloS one*, 9(4):e95053, 2014.

Robert J Elshire, Jeffrey C Glaubitz, Qi Sun, Jesse A Poland, Ken Kawamoto, Edward S Buckler, and Sharon E Mitchell. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, 6(5):e19379, 2011.

Patric Stephane Epopa, Abdoul Azize Millogo, Catherine Matilda Collins, Ace North, Frederic Tripet, Mark Quentin Benedict, and Abdoulaye Diabate. The use of sequential mark-release-recapture experiments to estimate population size, survival and dispersal of male mosquitoes of the anopheles gambiae complex in bana, a west african humid savannah village. *Parasites & vectors*, 10(1):1–15, 2017.

Sarah Evans, Jennifer BH Martiny, and Steven D Allison. Effects of dispersal and selection on stochastic assembly in microbial communities. *The ISME journal*, 11(1):176–185, 2017.

Igor Filipović, Hapuarachchige Chanditha Hapuarachchi, Wei-Ping Tien, Muhammad Aliff Bin Abdul Razak, Caleb Lee, Cheong Huat Tan, Gregor J Devine, and Gordana Rašić. Using spatial genetics to quantify mosquito dispersal for control programs. *BMC biology*, 18(1):1–15, 2020.

Lex Flagel, Yaniv Brandvain, and Daniel R Schrider. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular biology and evolution*, 36(2):220–238, 2019.

Maëva Gabrielli, Benoit Nabholz, Thibault Leroy, Borja Milá, and Christophe Thébaud. Within-island diversification in a passerine bird. *Proceedings of the Royal Society B*, 287(1923):20192999, 2020.

Madhav Gadgil. Dispersal: population consequences and evolution. *Ecology*, 52(2):253–261, 1971.

Graham Gower, Pablo Iáñez Picazo, Matteo Fumagalli, and Fernando Racimo. Detecting adaptive introgression in human evolution using convolutional neural networks. *Elife*, 10:e64669, 2021.

Adam S Hadley and Matthew G Betts. The effects of landscape fragmentation on pollination dynamics: absence of evidence not evidence of absence. *Biological Reviews*, 87(3):526–544, 2012.

Benjamin C Haller and Philipp W Messer. Slim 3: forward genetic simulations beyond the Wright-Fisher model. *Molecular biology and evolution*, 36(3):632–637, 2019.

Catriona M Harris, Kirsty J Park, Rachel Atkinson, Colin Edwards, and Justin MJ Travis. Invasive species control: incorporating demographic data and seed dispersal into a management model for *Rhododendron ponticum*. *Ecological Informatics*, 4(4):226–233, 2009.

Diana L Huestis, Adama Dao, Moussa Diallo, Zana L Sanogo, Djibril Samake, Alpha S Yaro, Yossi Ousman, Yvonne-Marie Linton, Asha Krishna, Laura Veru, et al. Windborne long-distance migration of malaria mosquitoes in the Sahel. *Nature*, 574(7778):404–408, 2019.

Jason M Jackson, Meaghan L Pimsler, Kennan Jeannet Oyen, Jonathan B Koch-Uhuad, James D Herndon, James P Strange, Michael E Dillon, and Jeffrey D Lozier. Distance, elevation and environment as drivers of diversity and divergence in bumble bees across latitude and altitude. *Molecular Ecology*, 27(14):2926–2942, 2018.

Joseph J Jacquot and Stephen H Vessey. Influence of the natal environment on dispersal of white-footed mice. *Behavioral Ecology and Sociobiology*, 37(6):407–412, 1995.

Moshe Jasper, Thomas L Schmidt, Nazni W Ahmad, Steven P Sinkins, and Ary A Hoffmann. A genomic approach to inferring kinship reveals limited intergenerational dispersal in the yellow fever mosquito. *Molecular Ecology Resources*, 19(5):1254–1264, 2019.

Michael D Jimenez, Edward E Bangs, Diane K Boyd, Douglas W Smith, Scott A Becker, David E Ausband, Susannah P Woodruff, Elizabeth H Bradley, Jim Holyan, and Kent Laudon. Wolf dispersal in the Rocky Mountains, Western United States: 1993–2008. *The Journal of Wildlife Management*, 81(4):581–592, 2017.

Joachim W Kadereit, Rami Arafeh, Gabriella Somogyi, and Erik Westberg. Terrestrial growth and marine dispersal? Comparative phylogeography of five coastal plant species at a European scale. *Taxon*, 54(4): 861–876, 2005.

Brian Keane. Dispersal and inbreeding avoidance in the white-footed mouse, *Peromyscus leucopus*. *Animal Behaviour*, 40(1):143–152, 1990.

Jerome Kelleher and Konrad Lohse. Coalescent simulation with msprime. In *Statistical Population Genomics*, pages 191–230. Humana, New York, NY, 2020.

Andrew D Kern and Daniel R Schrider. diploS/HIC: an updated approach to classifying selective sweeps. *G3: Genes, Genomes, Genetics*, 8(6):1959–1970, 2018.

Tony Kess, Anthony L Einfeldt, Brendan Wringe, Sarah J Lehnert, Kara KS Layton, Meghan C McBride, Dominique Robert, Jonathan Fisher, Arnault Le Bris, Cornelia den Heyer, et al. A putative structural variant and environmental variation associated with genomic divergence across the Northwest Atlantic in Atlantic Halibut. *ICES Journal of Marine Science*, 78(7):2371–2384, 2021.

Motoo Kimura and George H Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561, 1964.

Richard T Kingsford, James EM Watson, Carolyn J Lundquist, Oscar Venter, Lesley Hughes, EL Johnston, James Atherton, Mike Gawel, David A Keith, Brendan G Mackey, et al. Major conservation policy issues for biodiversity in Oceania. *Conservation Biology*, 23(4):834–840, 2009.

Ilpo Kojola, Jouni Aspi, Antero Hakala, Samuli Heikkinen, Catrin Ilmoni, and Seppo Ronkainen. Dispersal in an expanding wolf population in Finland. *Journal of Mammalogy*, 87(2):281–286, 2006.

Raphael Leblois, François Rousset, and Arnaud Estoup. Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics*, 166(2):1081–1092, 2004.

Mathew A Leibold, Marcel Holyoak, Nicolas Mouquet, Priyanga Amarasekare, Jonathan M Chase, Martha F Hoopes, Robert D Holt, Jonathan B Shurin, Richard Law, David Tilman, et al. The metacommunity concept: a framework for multi-scale community ecology. *Ecology letters*, 7(7):601–613, 2004.

Olivier Lepais, BEN Darvill, Stéphanie O'connor, Juliet L Osborne, Roy A Sanderson, John Cussans, Louis Goffe, and Dave Goulson. Estimation of bumblebee queen dispersal distances using sibship reconstruction method. *Molecular Ecology*, 19(4):819–831, 2010.

Chang Liu, Crista Bank, Michael Kersula, Geoffrey W Cowles, Douglas R Zemeckis, Steven X Cadrin, and Christopher McGuire. Movements of Atlantic halibut in the Gulf of Maine based on geolocation. *ICES Journal of Marine Science*, 76(7):2020–2032, 2019.

Neil F Lobo, Djibril M Sangaré, Allison A Regier, Kyanne R Reidenbach, David A Bretz, Maria V Sharakhova, Scott J Emrich, Sekou F Traore, Carlo Costantini, Nora J Besansky, et al. Breakpoint structure of the *Anopheles gambiae* 2Rb chromosomal inversion. *Malaria journal*, 9(1):1–9, 2010.

Erik Lundgren and Peter L Ralph. Are populations like a circuit? Comparing isolation by resistance to a new coalescent-based method. *Molecular ecology resources*, 19(6):1388–1406, 2019.

Joseph Marcus, Wooseok Ha, Rina Foygel Barber, and John Novembre. Fast and flexible estimation of effective migration surfaces. *Elife*, 10:e61927, 2021.

Jason Munshi-South, Christine P Zolnik, and Stephen E Harris. Population genomics of the Anthropocene: Urbanization is negatively associated with genome-wide variation in white-footed mouse populations. *Evolutionary applications*, 9(4):546–564, 2016.

Olaus J Murie and Adolph Murie. Travels of *Peromyscus*. *Journal of Mammalogy*, 12(3):200–209, 1931.

Ace North, Austin Burt, and H Charles J Godfray. Modelling the spatial spread of a homing endonuclease gene in a mosquito population. *Journal of Applied Ecology*, 50(5):1216–1225, 2013.

Ace R North, Austin Burt, and H Charles J Godfray. Modelling the potential of genetic control of malaria mosquitoes at national scale. *BMC biology*, 17(1):1–12, 2019.

Ace R North, Austin Burt, and H Charles J Godfray. Modelling the suppression of a malaria vector using a CRISPR-Cas9 gene drive to reduce female fertility. *BMC biology*, 18(1):1–14, 2020.

Reed F Noss, Erica Fleishman, Dominick A Dellasala, John M Fitzgerald, Mart R Gross, Martin B Main, Fiona Nagle, Stacey L O'Malley, and Jon Rosales. Priorities for improving the scientific foundation of conservation policy in North America. *Conservation Biology*, 23(4):825–833, 2009.

James Orsborne, Luis Furuya-Kanamori, Claire L Jeffries, Mojca Kristan, Abdul Rahim Mohammed, Yaw A Afrane, Kathleen O'Reilly, Eduardo Massad, Chris Drakeley, Thomas Walker, et al. Investigating the blood-host plasticity and dispersal of *Anopheles coluzzii* using a novel field-based methodology. *Parasites & vectors*, 12(1):1–8, 2019.

Juliet L Osborne, Andrew P Martin, Norman L Carreck, Jennifer L Swain, Mairi E Knight, Dave Goulson, Roddy J Hale, and Roy A Sanderson. Bumblebee flight distances in relation to the forage landscape. *Journal of animal ecology*, 77(2):406–415, 2008.

Emmanuel Paradis, Stephen R Baillie, William J Sutherland, and Richard D Gregory. Patterns of natal and breeding dispersal in birds. *Journal of Animal ecology*, 67(4):518–536, 1998.

Rémy S Pasquet, Alexis Peltier, Matthew B Hufford, Emeline Oudin, Jonathan Saulnier, Lénaic Paul, Jette T Knudsen, Hans R Herren, and Paul Gepts. Long-distance pollen flow assessment through evaluation of pollinator foraging range suggests transgene escape distances. *Proceedings of the National Academy of Sciences*, 105(36):13456–13461, 2008.

Desislava Petkova, John Novembre, and Matthew Stephens. Visualizing spatial population structure with estimated effective migration surfaces. *Nature genetics*, 48(1):94–100, 2016.

Andrew S Pullin, Andras Baldi, Ozgun Emre Can, Martin Dieterich, Vassiliki Kati, Barbara Livoreil, Gabor Lövei, Barbara Mihok, Owen Nevin, Nuria Selva, et al. Conservation focus on europe: major conservation policy issues that need to be informed by conservation science. *Conservation biology*, 23(4):818–824, 2009.

Michelle M Riehle, Tullu Bukhari, Awa Gneme, Wamdaogo M Guelbeogo, Boubacar Coulibaly, Abdrahamane Fofana, Adrien Pain, Emmanuel Bischoff, Francois Renaud, Abdoul H Beavogui, et al. The *Anopheles gambiae* 2La chromosome inversion is associated with susceptibility to *Plasmodium falciparum* in Africa. *Elife*, 6:e25813, 2017.

Harald Ringbauer, Graham Coop, and Nicholas H Barton. Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351, 2017.

Colin G Rose, Kennedy T Paynter, and Matthew P Hare. Isolation by distance in the eastern oyster, *Crassostrea virginica*, in Chesapeake Bay. *Journal of Heredity*, 97(2):158–170, 2006.

F Rousset. Genetic differentiation between individuals. *J Evol Biol*, 13:58–62, 2000.

François Rousset. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145(4):1219–1228, 1997.

Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660, 2021.

Daniel R Schrider and Andrew D Kern. Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, 34(4):301–312, 2018.

Rena M Schweizer, Bridgett M Vonholdt, Ryan Harrigan, James C Knowles, Marco Musiani, David Coltman, John Novembre, and Robert K Wayne. Genetic subdivision and candidate genes under selection in North American grey wolves. *Molecular ecology*, 25(1):380–402, 2016.

Nancy L Shackell, Jonathan AD Fisher, Cornelia E den Heyer, Daniel R Hennen, Andrew C Seitz, Arnault Le Bris, Dominique Robert, Michael E Kersula, Steven X Cadrin, Richard S McBride, et al. Spatial ecology of Atlantic Halibut across the Northwest Atlantic: a recovering species in an era of climate change. *Reviews in Fisheries Science & Aquaculture*, pages 1–25, 2021.

Sara Sheehan and Yun S Song. Deep learning for population genetic inference. *PLoS computational biology*, 12(3):e1004845, 2016.

Ashlee Shipham, Daniel J Schmidt, and Jane M Hughes. Indirect estimates of natal dispersal distance from genetic data in a stream-dwelling fish (*Mogurnda adspersa*). *Journal of Heredity*, 104(6):779–790, 2013.

William J Sutherland, Susan Armstrong-Brown, Paul R Armsworth, Brereton Tom, Jonathan Brickland, Colin D Campbell, Daniel E Chamberlain, Andrew I Cooke, Nicholas K Dulvy, Nicholas R Dusic, et al. The identification of 100 ecological questions of high policy relevance in the UK. *Journal of applied ecology*, 43(4):617–627, 2006.

Marco Todesco, Gregory L Owens, Natalia Bercovich, Jean-Sébastien Légaré, Shaghayegh Soudi, Dylan O Burge, Kaichi Huang, Katherine L Ostevik, Emily Drummond, Ivana Imerovski, et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, 584(7822):602–607, 2020.

Justin MJ Travis, Maria Delgado, Greta Bocedi, Michel Baguette, Kamil Bartoń, Dries Bonte, Isabelle Boulangeat, Jenny A Hodgson, Alexander Kubisch, Vincenzo Penteriani, et al. Dispersal and species' responses to climate change. *Oikos*, 122(11):1532–1540, 2013.

Antariksh Tyagi, Shivani Singh, Parneeta Mishra, Akanksha Singh, Abhinandan Mani Tripathi, Satya Narayan Jena, and Sribash Roy. Genetic diversity and population structure of *Arabidopsis thaliana* along an altitudinal gradient. *AoB Plants*, 8, 2016.

Bénédikte Vercaemer, Philippe St-Onge, Koren Spence, Shauna Gould, and Allison McIsaac. Assessment of biodiversity of American oyster (*Crassostrea virginica*) populations of Cape Breton, NS and the Maritimes. *Canadian Technical Report of Fisheries and Aquatic Sciences*, 2872, 2010.

P Kirk Visscher and Thomas D Seeley. Foraging strategy of honeybee colonies in a temperate deciduous forest. *Ecology*, 63(6):1790–1801, 1982.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Naomi J Wender, Chandra R Polisetty, and Kathleen Donohue. Density-dependent processes influencing the evolutionary dynamics of dispersal: a functional analysis of seed dispersal in *Arabidopsis thaliana* (Brassicaceae). *American Journal of Botany*, 92(6):960–971, 2005.

John A Wiens, Diana Stralberg, Dennis Jongsomjit, Christine A Howell, and Mark A Snyder. Niches, models, and climate change: assessing the assumptions and uncertainties. *Proceedings of the National Academy of Sciences*, 106(Supplement 2):19729–19736, 2009.

John J Wiens. Climate-related local extinctions are already widespread among plant and animal species. *PLoS biology*, 14(12):e2001104, 2016.

Sewall Wright. Isolation by distance. *Genetics*, 28(2):114, 1943.

Sewall Wright. Isolation by distance under diverse systems of mating. *Genetics*, 31(1):39, 1946.

Supplementary material

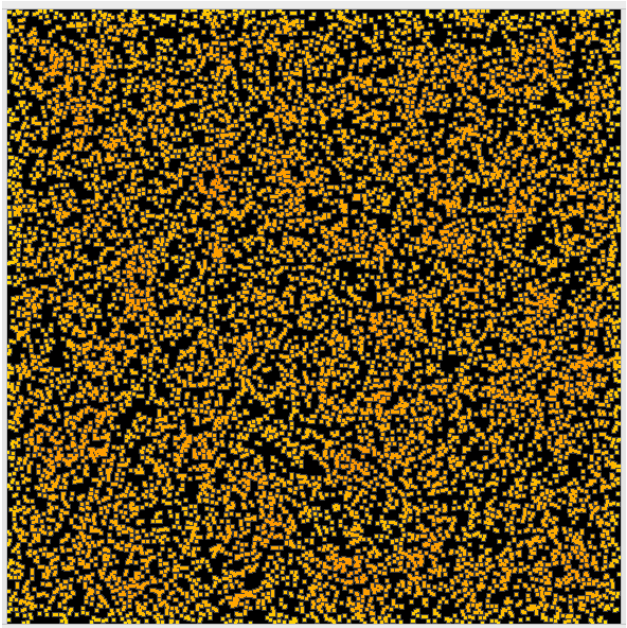

Figure S1: Screenshot of an example simulation in SLiM's graphical user interface. Parameter Set 1 with $\sigma = 1.0$.

.

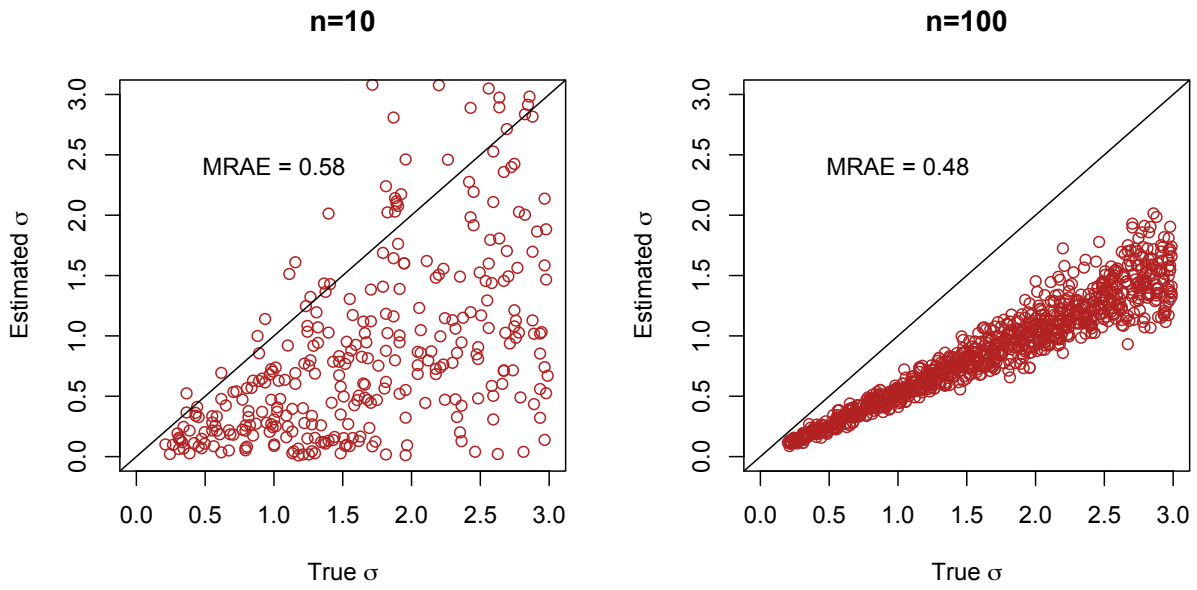

Figure S2: Predictions on simulated data using `IBD-Analysis` with identity-by-descent blocks empirically derived from the `Refined IBD` program (Browning and Browning, 2013). With $n = 10$, zero identity-by-descent blocks were detected in 4% of datasets, and `IBD-Analysis` did not converge for an additional 63% of datasets. The mean RAE from $n = 10$ using inferred identity-by-descent blocks was lower than using perfect identity-by-descent blocks due to fewer extreme outliers; the median RAE was 0.4 with perfect identity-by-descent blocks, and 0.58 with inferred identity-by-descent blocks.

.

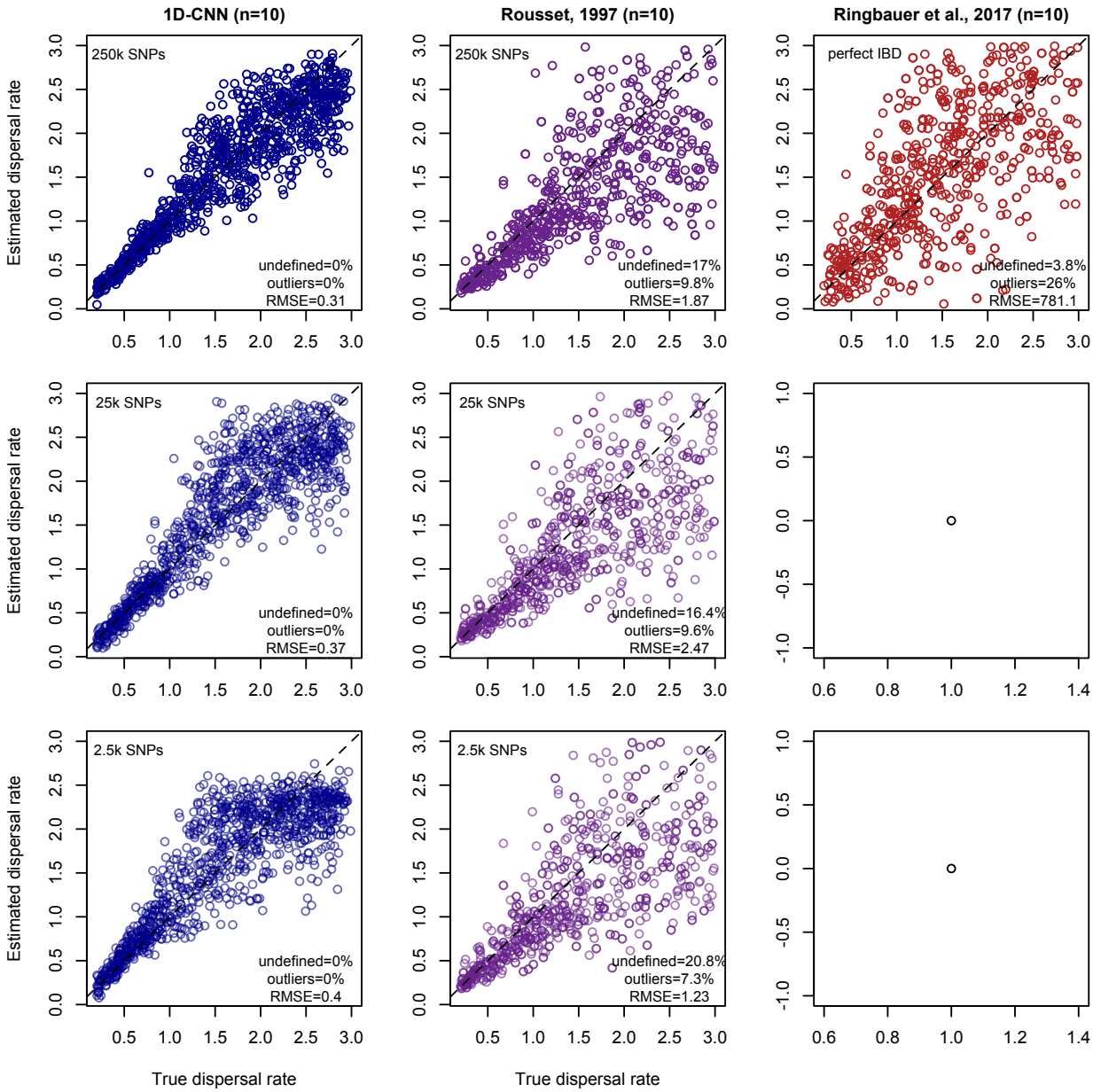

Figure S3: Comparison with other methods, $n = 10$, and varying SNP number (other parameters as in Parameter Set 1). The `IBD-Analysis` plot used perfect identity-by-descent tracts rather than SNPs.

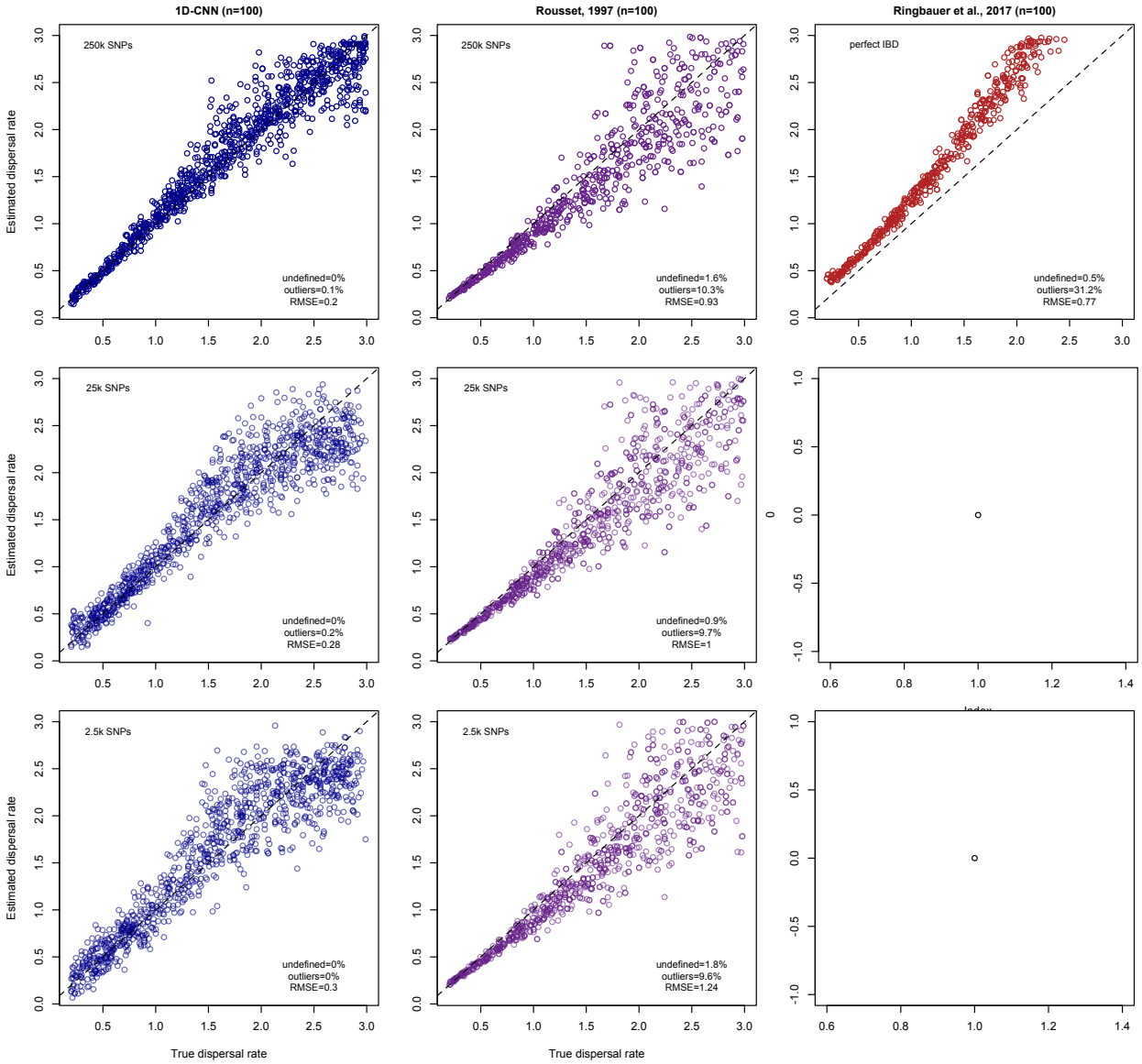

Figure S4: Comparison with with other methods, $n = 100$, and varying SNP number (other parameters as in Parameter Set 1). The `IBD-Analysis` plot used perfect identity-by-descent tracts rather than SNPs.

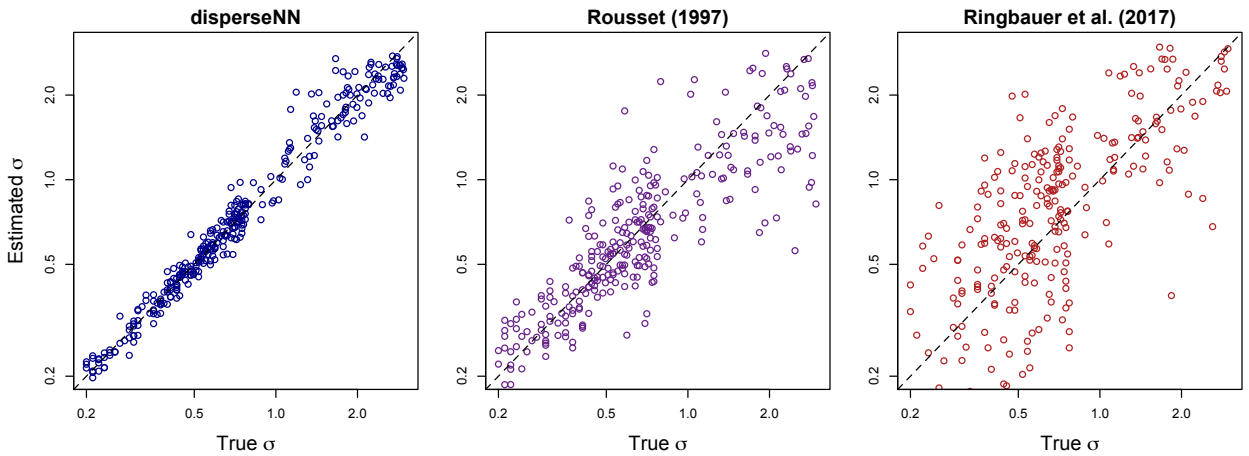

Figure S5: Predictions with log-transformation to show relative error, $n = 10$ (Parameter Set 1). Data points in the larger half of the $\log(\sigma)$ range were down-sampled to one-half the number of points in the smaller half of the range to obtain roughly even density of points across the range of $\log(\sigma)$. Before down-sampling, points were more dense towards the right-hand side.

.

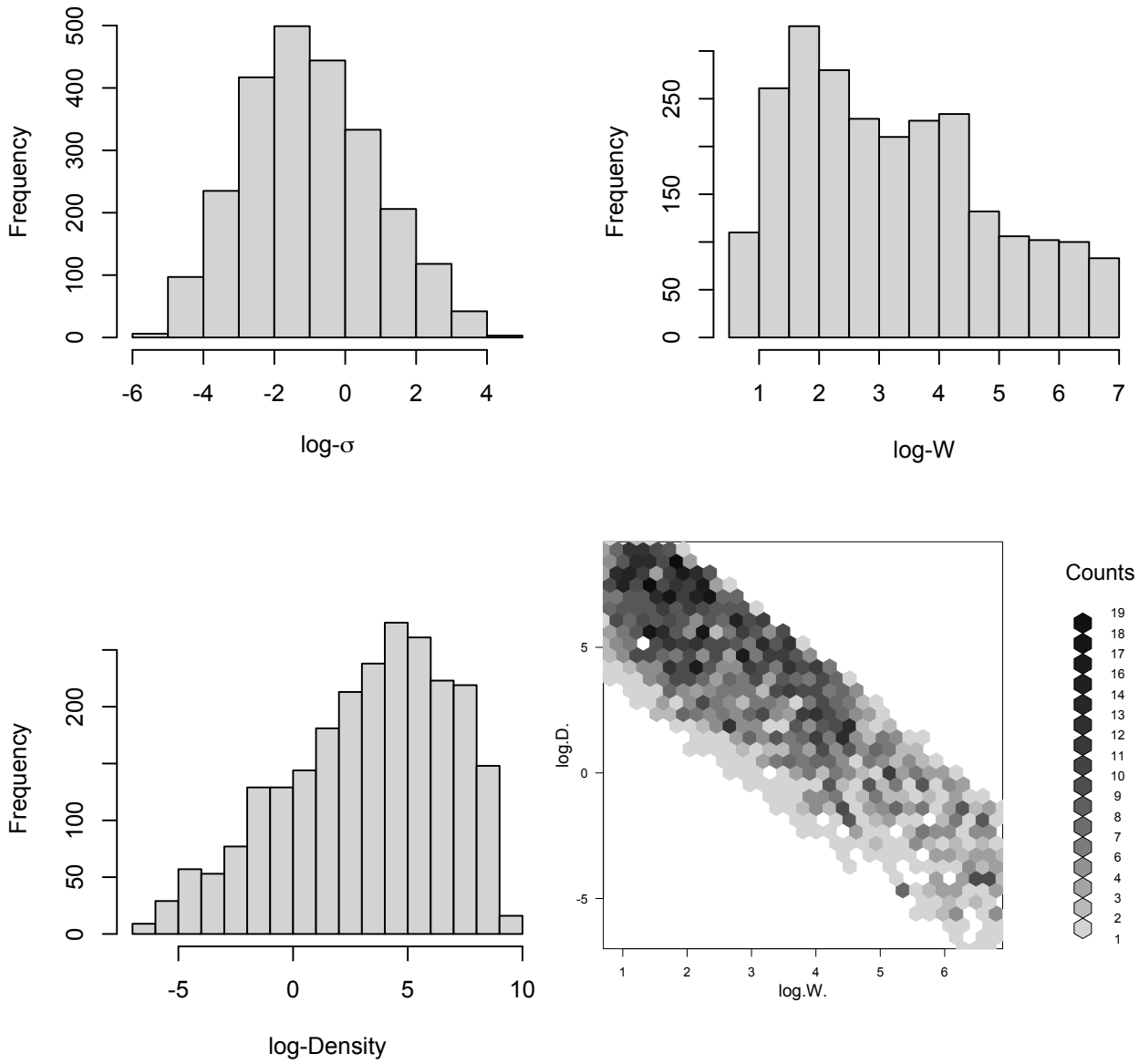

Figure S6: Realized training distributions for empirical analysis (Parameter Set 11). "W" is habitat width. Some areas of parameter space could not be simulated due to population extinction or computational limitations.

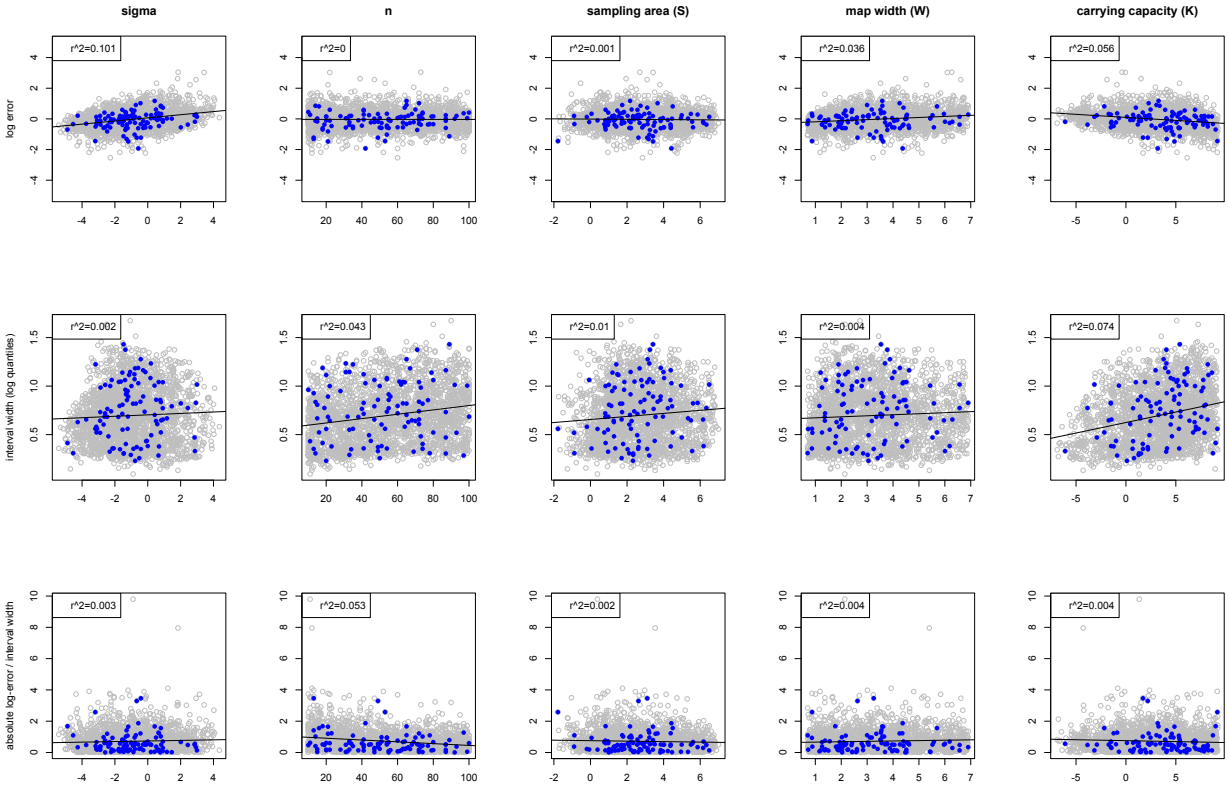

Figure S7: Exploring the effects of five different predictor variables—(1) σ, (2) n, (3) sampling area, (4) map width, (5) carrying capacity—on three different response variables (A) log error, (B) the interval width of the log-transformed middle 95% range of the bootstrap distribution, and (C) absolute log-error divided by the interval width (Parameter Set 11). Shown are 2400 datasets including both held-out test data (blue; 100 datasets) and training data (grey; 2,300 datasets). The line of best fit and r^2 include all 2400 data points.

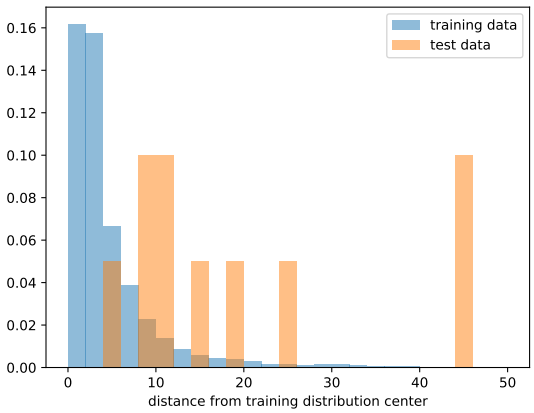

Figure S8: Mahalonobis distance from the center of the training distribution with respect to five summary statistics: nucleotide diversity, Tajima's D, inbreeding coefficient, observed heterozygosity, and expected heterozygosity (Parameter Set 11). "test data" are the empirical datasets.

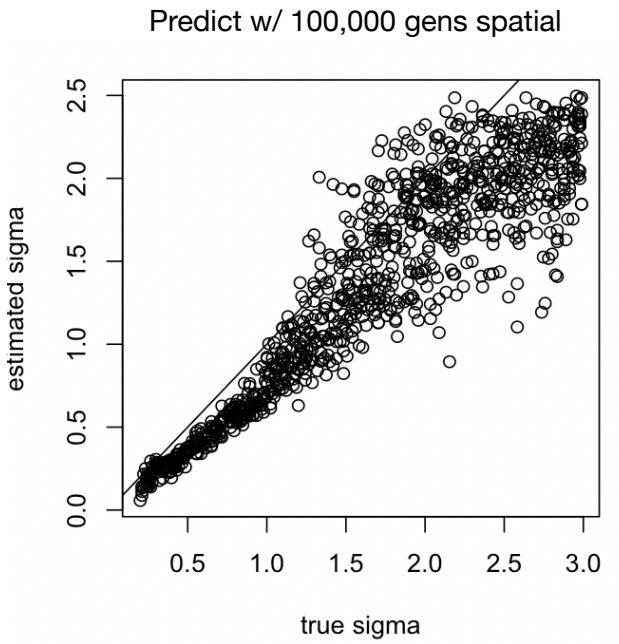

Figure S9: CNN trained with only 100 generations in spatial SLiM before recapitation with msprime (other parameters as in Parameter Set 2). The depicted results are from testing on simulations with 100,000 generations spatial, which is nearly full-spatial.

.

Appendix: stuff we tried that did not work

This Appendix describes analyses not included in the main document, including strategies that didn't work.

A1. Attempts to use sampling localities

It is intuitive that signal about dispersal might be gleaned from the individual sample locations, as previous population-genetics-based inference methods use sample locations as input. We tried the following strategies for showing the sample locations to the CNN. In each experiment, we modified the neural network architecture to accommodate the sample locations in various ways. Otherwise, the neural network in each experiment closely resembled the architecture described in the main text.

- *Table of locations.* An $n \times 2$ array containing the x and y coordinates was shown to the CNN in a separate input branch (in place of the sampling width input). This input went through a single 128-unit dense layer with ReLu activation before flattening and concatenating with the previous branch.
- *Stored in genotype matrix* Additional rows in the genotype matrix were used to store the x and y coordinates for each individual.
- *3-channel array.* A 3-dimensional array was used to store (1) the genotypes, (2) x coordinates, and (3) y coordinates. In the second and third channels, the spatial coordinates were repeated for m rows equal to the number of SNPs. Here, the neural network used 1D-convolution and pooling layers, as described in the main text, however the convolution and pooling layers spanned all three channels simultaneously.
- *2D CNN.* We also tried a variation of the the 3-channel-array strategy using 2D-convolution and pooling layers with a 2x2 window.

For each of the above strategies, we trained the neural network in the same manner as the "baseline" model from the misspecification analysis in the main text. The outcome for each was the same: the mean RAE was indistinguishable from the baseline model that does not include sample locations. Moreover, we shuffled the sample locations input, such that each individual has a randomly assigned location, and the output was unchanged. Our interpretation is that the CNN ignores the location data in the experiments attempted thus far, either because the locations are not necessary for estimating σ, or because we failed to effectively show the network the locations.

A2. Including isolation-by-distance summary statistics

We tested whether isolation by distance information in the form of summary statistics would improve inference of σ. Specifically, we summarized isolation-by-distance as:

- b, the slope of the line of best fit to genetic distances versus geographic distances.
- r^2, the coefficient of correlation between genetic distance and geographic distance.

Including either (or both) of these statistics as a separate input branch of size one (or two) marginally improved validation accuracy. The new input branch went through a 128-unit dense layer with ReLu activation before concatenating with the previous branch. Thus, future empirical applications might explore using the above or different summary statistics alongside the genotype matrix for estimating σ, or other population genetic parameters. We did not present these results in the main text because (1) the benefit was negligible, and (2) it is beyond the scope of our study to decide on the most relevant and appropriate summary statistics, as countless other statistics might be evaluated for use with, or without, the genotype matrix that we used.