Genetic assignment of individuals to source populations using network estimation tools

Markku Kuismin[1,2] Dilan Saatoglu[3] Alina K. Niskanen[3,4] Henrik Jensen[3], and Mikko J. Sillanpää*[1,2,5]

[1]*Research Unit of Mathematical Sciences, University of Oulu, Finland*
[2]*Biocenter Oulu, University of Oulu, Finland*
[3]*Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, Norway*
[4]*Ecology and Genetics Research Unit, University of Oulu, Finland*
[5]*Infotech Oulu, University of Oulu, Finland*

October 24, 2019

Abstract

1. Dispersal, the movement of individuals between populations, is crucial in many ecological and genetic processes. However, direct identification of dispersing individuals is difficult or impossible in natural populations. By using genetic assignment methods, individuals with unknown genetic origin can be assigned to source populations. This knowledge is necessary in studying many key questions in ecology, evolution and conservation.
2. We introduce a network-based tool BONE (Baseline Oriented Network Estimation) for genetic population assignment, which borrows concepts from undirected graph inference. In particular, we use sparse multinomial Least Absolute Shrinkage and Selection Operator (LASSO) regression to estimate probability of the origin of all mixture individuals and their mixture proportions without tedious selection of the LASSO tuning parameter. We compare BONE with three genetic assignment methods implemented in R packages `radmixture`, `assignPOP` and `RUBIAS`.
3. Probability of the origin and mixture proportion estimates of both simulated and real data (an insular house sparrow metapopulation and Chinook salmon populations) given by BONE are competitive or superior compared to other assignment methods. Our examples illustrate how the network estimation method adapts to population assignment, combining the efficiency and attractive properties of sparse network representation and model selection properties of the L_1

*Corresponding author: mikko.sillanpaa@oulu.fi

regularization. As far as we know, this is the first approach showing how one can use network tools for genetic identification of individuals' source populations.

4. BONE is aimed at any researcher performing genetic assignment and trying to infer the genetic population structure. Compared to other methods, our approach also identifies outlying mixture individuals that could originate outside of the baseline populations. BONE is a freely available R package under the GPL license and can be downloaded at GitHub. In addition to the R package, a tutorial for BONE is available at https://github.com/markkukuismin/BONE/.

Keywords: assignment analysis, genetic stock identification, LASSO, networks, SNP

1 Introduction

Dispersal is of fundamental importance in ecology, evolutionary biology, conservation and management (see, e.g., Ronce, 2007; Clobert et al., 2012; Driscoll et al., 2014; Saastamoinen et al., 2018). Being able to accurately identify dispersers and individuals' population of origin is therefore important. One fruitful approach is to assign individuals to their population of origin using genetic information (Manel et al., 2005). Multilocus genetic data, such as single-nucleotide polymorphisms (SNP), are now a common source of information, for example, in genetic stock identification (GSI) in fishery management (see, e.g., Beacham et al., 2012; Garvin et al., 2010). This is due to the major development in sequencing technologies (review in Jiang et al., 2016; Garvin et al., 2010). Today it is increasingly common to have data sets with thousands of markers on a variety of species and thus new methods are emerging to analyze large data sets (Li et al., 2008; Novembre et al., 2008; Anderson et al., 2008; Anderson, 2010; Helyar et al., 2011; Ruegg et al., 2017; McKinney et al., 2017).

Utilizing graphs in computational biology has been an active endeavour for the past couple of decades (review in Wang and Huang, 2014). These methods include directed and undirected graphical models, which characterize the conditional dependency structure between random variables (review in Drton and Maathuis, 2017). One widely used variant is the weighted gene co-expression network analysis (see, e.g., Horvath, 2011). Graph methods are not only restricted to gene networks but they have also been utilized in landscape genetics (see, e.g., Garroway et al., 2008), phylogenetic trees construction (see, e.g., Huson and Scornavacca, 2011) and in estimation of genetic population structure (see, e.g., Dyer and Nason, 2004; Greenbaum et al., 2016; Kuismin et al., 2017), to mention a few applications.

Here we show how a graph estimated with the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) can be applied for population assignment using multilocus SNP data. We have used a similar approach previously for genetic population structure inference (Kuismin et al., 2017). The difference of genetic assignment compared to the population structure estimation is that some information on populations of origin (defined as "baseline" or "source populations") is known in advance, and these baseline

populations function as reference data to which individuals without known origin (defined as "mixture individuals", included in a "mixture population") are assigned.

We show how the probability of the origin and the mixing proportions of an individual can be determined from a graph. In our method, these probabilities are estimated based on the robustness of the results when the graph is produced using different values of the LASSO tuning parameter that controls the sparsity level of the graph. We can inspect the strength of the probability of the origin by examining how graph neighborhood of a mixture individual changes with different regularizations along the LASSO solution path. We claim that if the mixture individual has a strong genetic similarity with individuals in a baseline population, the neighborhood of such a mixture individual is more robust to the changes of the LASSO tuning parameter value compared to those mixture individuals that have a weak genetic similarity with baseline individuals. In particular, with strong penalization, the strong genetic signals stand out more clearly than the weaker ones. If neighbors of an individual seem to resurface more or less randomly from the different baseline populations when the LASSO regularization is changed, one can suspect that this individual is not a member of any of the baseline populations. Hereafter, we refer this kind of individuals to as "outsiders". Here, we first describe in detail our network-based method.

We can describe the workflow of our method with five steps:

1. Use multinomial LASSO regression to determine close genetic relationships between individuals in baseline populations, with known structure, and mixture individuals, with unknown origin. By "close genetic relationship" we mean that both the mixture individual and baseline individual belong to the same group (baseline population) that is genetically homogeneous.

2. Collect results of the multinomial LASSO regression into a form of a genetic relationship network.

3. Assign mixture individuals to those baseline population(s) with which the node of the mixture individual shares at least one edge (neighbor) in the network.

4. Determine estimates of the probability of the origin based on the node degree and mixture proportions, which are averaged probabilities of the origins over different

baseline populations.

5. Detect which mixture individuals are potential outsiders using the whole LASSO solution path information.

We call our method BONE, that stands for Baseline Oriented Network Estimation.

Next, we compare BONE with three population assignment methods: commonly used ADMIXTURE (Alexander et al., 2009), along with two recently published methods "RUBIAS" (Moran and Anderson, 2019) and "AssignPOP" (Chen et al., 2018) using both simulated and real data. ADMIXTURE, RUBIAS and AssignPOP are implemented in R packages `radmixture`, `RUBIAS` and `AssignPOP`, respectively, and are publicly available at CRAN (The Comprehensive R Archive Network). RUBIAS and AssignPOP in particular are designed for population assignment. They allow assignment of individuals in mixed populations with several classification tools (support vector machines, naive Bayes etc.), they also enable simulation of mixture individuals to predict assignment accuracy. ADMIXTURE is usually used to estimate ancestries when nothing is known about the contributing ancestral populations, but it can also be applied to population assignment using so called supervised learning model extension (Alexander and Lange, 2011).

2 Methods

2.1 Baseline-mixture graph

Assignments of mixture individuals, with unknown genetic background, into known baseline populations, can be represented with baseline-mixture graph (terms graph and network are used interchangeably in this article). Let M denote the group of mixture individuals and B the group of baseline individuals (group of source populations or reporting units). There are no duplicates in the data meaning that sets M and B are separated, $M \cap B = \emptyset$. We refer to different baseline populations with subscripts B_{bp}. Assume that there are $|B| = p$ baseline individuals and $|M| = m$ mixture individuals. Notation $|\cdot|$ is the cardinality of the set (i.e., how many individuals there are).

Let $G = (E, V)$ be an undirected graph (the baseline-mixture graph). The set $E =$

$\{i, j \mid i \in M, j \in B\}$ is the set of nodes. Each node i corresponds to a mixture individual and each node j corresponds to a baseline individual. $V = \{(i,j),(j,i) \mid i \in M, j \in B\}$ is the set of edges. We shortly write $(i,j) \in V$, denoting, that both pairs (i,j) and (j,i) are in the set V. If $(i,j) \in V$, then there is an undirected edge between nodes i and j in the graph. Undirected edge (i,j) denotes the close genetic relationship between mixture individual i and baseline individual j and vice versa. The undirected graph does not imply direction of the genetic relationship.

The undirected graph G can be coded into a symmetric, $(p+m)\times(p+m)$ binary valued adjacency matrix $A = [a_{ij}]$, $a_{ij} \in \{0,1\}$. The diagonal elements of the adjacency matrix are all zero, because there are no edges from a node to itself (loops) in the undirected graph G, $\mathrm{diag}(A) = 0$. We define $a_{ij} = a_{ji} = 1$, if $(i,j) \in V$, when there is a close genetic relationship between mixture individual i and baseline individual j, and $a_{ij} = a_{ji} = 0$, if $(i,j) \notin V$, when there is no genetic relationship between mixture individual i and baseline individual j. Because of this, the adjacency matrix A is symmetric, $A = A^\top$, where superscript $^\top$ denotes matrix or vector transposition.

Neighbors of the node $i \in M$ are all nodes $j \in B$ that share an edge with the node $i \in M$. Together, neighbors form a neighborhood of node i. In the baseline-mixture graph, we require that mixture individuals cannot be neighbors of other mixture individuals (no loops). Similar restriction holds also for baseline individuals. We have illustrated this in Fig 1.

2.2 Probability of the origin and mixing proportions

We use the symmetric adjacency matrix A to describe how one can determine the probability of the origin and mixture proportions of each mixture individual. All we need is the node degrees and the number of neighbors of each mixture individual. Denote the probability of the origin of a mixture individual i with $(b+1)\times 1$ vector $p(Mix_i \mid B)$ where $b = \#\{\text{baseline populations}\}$ (note that $b = 3$ in the illustrative example of Fig 1). Vector element $p(Mix_i \mid B_{bp})$ is the probability of the origin of individual i originating from the baseline population B_{bp}, $bp = \{1, \ldots b\}$. Populations are not in any particular order. We define the probability of the origin $p(Mix_i \mid B_{bp})$ as follows:

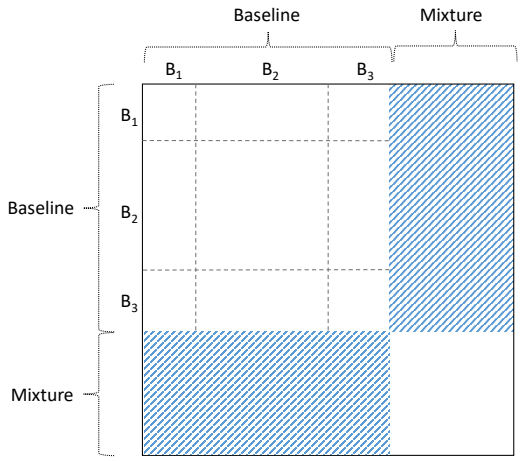

Figure 1: A schematic illustration of the symmetric adjacency matrix. Assignments of the baseline source individuals are known and fixed into one of the baseline source populations before analysis. In this illustrative example, there are three baseline source populations: B1, B2 and B3. Potential neighborhoods that are estimated using BONE correspond to the shaded areas, and form the fundamental set for neighborhood selection of the baseline-mixture graph.

$$p(Mix_i \mid B_{bp}) = \frac{1}{d_i} \sum_{j \in B_{bp}} a_{ij}, \tag{1}$$

where $\sum_{j \in B_{bp}} a_{ij}$ is the number of neighbours (baseline individuals) of individual i that belong to population B_{bp} and $d_i = \sum_{k=1}^{p} a_{ik}$, $i \in \{(p+1), \ldots (p+m)\}$, is the number of neighbors of individual i. Here d_i can have values between zero and p. It is possible that an individual is originating from outside of the baseline populations. In this case, $d_i = 0$ and we set $p(Mix_i \mid B_{(b+1)}) = 1$, which denotes an outsider population.

Mixing proportion of population B_{bp} is the average of the probability of the origin estimates defined in equation (1) over all mixture individuals assigned in the baseline population B_{bp}

$$p(B_{bp}) = \frac{1}{m} \sum_{i=1}^{m} p(Mix_i \mid B_{bp}), \tag{2}$$

where m is the total number of mixture individuals. In the next section, we describe how one can estimate the baseline-mixture graph G needed to estimate the probability of the origin and mixture proportions.

2.3 Network estimation with LASSO

It is well known that the LASSO method can be used in a sparse linear model selection due to the properties of the L_1-type penalty function (see, e.g., Tibshirani, 1996; Meinshausen and Bühlmann, 2006). Following the work of Kuismin et al. (2017), we use the L_1-regularized multinomial logit model (Friedman et al., 2010) to select the graph and estimate the probability of the origin for each mixture individual under the restrictions illustrated in Fig 1.

We have described the multinomial logit model used in our neighborhood selection in more detail in supplementary materials. We have also depicted other elements of the graph, probability of the origin, mixture proportions and quick decision rule for outsider detection based on this LASSO model in more detail in supplementary materials.

It is important for the user to know, that the sparsity of the network estimated with the LASSO regularized multinomial logit model is controlled in BONE with a user defined *tuning parameter* λ that is a non-negative real number. When λ is sufficiently large, the selected graph is sparse or even empty. When the tuning parameter value decreases, more edges are included into the selected network. The LASSO procedure is computationally very efficient and (sometimes) it is actually faster to compute LASSO solution using a decreasing sequence of tuning parameter values down to a small λ value, than computing LASSO solution only at one small value for the tuning parameter (Friedman et al., 2010). The information gained from this procedure is used to compute so-called solution path that is the set of LASSO estimates associated with each pre-determined value of the tuning parameter.

We propose two methods of selecting neighbors for mixture individuals from the baseline group:

1. **Winner Takes it All (WTA) method**: This method starts with an empty graph and follows the LASSO-solution path using a decreasing sequence of tuning parameter

values until at least one neighbor for the mixture individual is found from a baseline group. The mixture individual is assigned to the same baseline population where the first neighbor was found. Occasionally, it is possible to find more neighbors than just one with this procedure, which causes the mixture individual to be assigned to all of the baseline populations where the neighbours were found. This procedure is repeated for all mixture individuals, after which all found neighbors are summarized into an symmetric adjacency matrix as described before.

2. **Solution path method**: In this method, the whole LASSO-solution path is examined and the mixture individual is associated to all baseline population neighbors that are found at each tuning parameter value. In other words, a continuum of graphs is explored and all neighbors are collected over all neighborhoods found. Neighbors for all mixture individuals are summarized into a weighted graph: we assume that graphs composed with large tuning parameter values contribute more to the weighted graph and thus we give them larger weights.

The solution path method is loosely related to Lockhart et al. (2014) where the authors use LASSO-solution path in significance testing. The probability of the origin can again be computed from the composed adjacency matrix (1). We have described the computing of the weighted adjacency matrix, mentioned in the solution path method, in more detail in supplementary materials. Both the WTA and the solution path method ease the selection of the LASSO tuning parameter value, which otherwise is a laborious task (see, e.g., Meinshausen and Bühlmann, 2010): inclusion of a model parameter does not (excessively) complicate BONE. It is sufficient just to use a wide range of tuning parameter values from large to small ones.

The weakness of the multinomial logit model is that if the multilocus genotype pattern of some individual has extremely low occurrence of a certain genotype class, e.g., if an individual has very low heterozygosity, the numerical algorithm used to solve the multinomial logit LASSO problem returns biased estimates. The algorithm may fail in whole if there are one or less occurrences of a certain genotype class (either heterozygous or homozygous) or an individual is totally homozygous/heterozygous over loci. Nevertheless, this rarely happens if the data contains thousands or hundreds of thousands of SNPs. We have illustrated

this problem in supplementary Fig "BONEFailureSchematics".

2.4 Qualitative check of the randomness of the probability of the origin

If the neighborhoods of mixture individuals were defined by chance, then the observed count of the neighbors of a mixture individual would be close to the counts of the individuals in each baseline population at hand. In this case, the probability of a mixture individual to be a neighbor with a baseline individual in group B_{bp} is p_{bp}/p. Here p_{bp} is the number of individuals in baseline population B_{bp} and also the expected count of neighbors, whereas p is the total number of baseline individuals. If the observed size of the neighborhood is not considerably different from the expected counts, it is possible that the mixture individual is an outsider (e.g., a disperser from outside of the baseline populations).

However, the number of neighbors in the network is very limited: a given mixture individual usually has just a few or no neighbors in the baseline populations. This is because BONE produces very sparse network estimates which are easier to examine in general. Since the number of observed counts (number of neighbors) is very limited, statistical testing, just as G-test, of the neighborhood division is not feasible. Nevertheless, we propose to check how observed neighborhood proportions (probability of the origin estimates) diverge from expected neighborhood proportions, which are equal to the expected probability of origins.

To identify potential outsiders in the data, we compute the mean squared errors (MSE) between probability of the origins, estimated by using the solution path method, and expected probability of the origins. When the tuning parameter value is decreased, probability of the origin estimates approach the expected values. If the tuning parameter value is zero, all probability of the origin estimates are equal to the expected values because every baseline individual is a neighbor of each mixture individual (full graph). However, the LASSO augmented multinomial logit model first finds neighbors of a mixture individual from the baseline with strong genetic similarity (shared SNP genotype patterns) and these assignments are robust to the increment of the tuning parameter value by showing constantly strong degree of genetic similarity. For the individuals whose neighborhood is found only

with very small tuning parameter values, the probability of the origin estimates are very close to the expected probability of the origin values. Their neighborhood approaches the full graph and only the size of the baseline population determines their neighborhood, seemingly found by chance. Individuals whose probability of the origin estimates are the closest to the expected values are qualitatively interpreted as potential outsiders. In this article, we use the mean squared error to measure the distance between estimated and expected probability of the origin values. This deduction rule cannot be applied with the Winner Takes it All method, because it produces rigid estimates for the probability of the origin, which seemingly always differ from the expected probability of the origin. Mathematical basis of this qualitative procedure is described in more detail in supplementary materials.

Individuals identified with this procedure have fragmented probability of the origin estimates. This can be caused either by i) dispersal from outside of the baseline group, or ii) strong admixture of the baseline populations. We have illustrated different outcomes of BONE and what might cause these outcomes in Fig 2.

2.5 Comparison of different assignment methods using simulated and partitioned real data

We utilize three different data sets (a house sparrow data set, and a large and a small Chinook salmon data set) in two different illustrative examples, which we describe below.

2.5.1 Data simulation scheme

We simulate SNP genotype data using the Diriclet-multinomial model of the ADMIXTURE (Alexander et al., 2009) producing the binomial proportions

$$\begin{aligned}
\Pr(0 \text{ for } i \text{ at SNP } j) &= \left[\sum_k q_{ik} f_{kj}\right]^2, \\
\Pr(1 \text{ for } i \text{ at SNP } j) &= 2\left[\sum_k q_{ik} f_{kj}\right]\left[\sum_k q_{ik}(1-f_{kj})\right], \\
\Pr(2 \text{ for } i \text{ at SNP } j) &= \left[\sum_k q_{ik}(1-f_{kj})\right]^2,
\end{aligned} \tag{3}$$

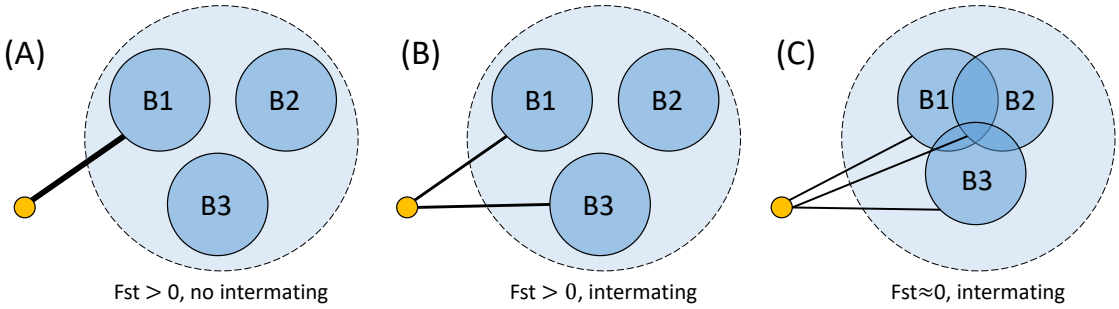

Figure 2: An illustration of how subpopulation division (here, B1, B2 and B3) of the baseline group (light shaded area) and intermating between baseline populations (admixture) may affect assignments of a mixture individual (the small orange node). The width of the edge represents the strength of the assignment. **(A)** Baseline populations are clearly separated and there is no admixture: assignments are clear and potential outsiders are easy to detect. **(B)** Baseline populations are clearly separated and the mixture individual has relatives in multiple baseline populations: assignments are more ambiguous and admixed individuals might be identified as potential outsiders. **(C)** Baseline populations are not clearly separated and there is intermating: assignments seem to happen by chance and outsider detection identifies many potential outsiders.

where $Q = [q_{ik}]$ is the $(I \times K)$ ancestry coefficient matrix and $F = [f_{kj}]$ is the $(K \times J)$ population allele frequency matrix. Here I is the number of unrelated individuals, K is the number of populations and J is the number of SNPs.

In our simulation study, Q is fixed such that individuals from different populations are totally separated from individuals from other populations (see Fig 2 (**A**)). We use empirical SNP genotype frequencies as known genotype frequencies. These genotype frequencies are determined from two data sets which we use as our starting point:

1. An extensive house sparrow data set (Lundregan et al., 2018; Araya-Ajoy et al., 2019; Saatoglu et al., 2019, see supplementary materials for details). Overall, there are 183,145 SNPs and 507 baseline birds for 2012 in the original data file from eight populations.

2. A large Chinook salmon data set with several thousands of SNPs (Larson et al., 2014).

This data set is freely available in Dryad digital repository (`https://doi.org/10.5061/dryad.rs4v1`). Overall, there are 10,944 SNPs and 265 individuals for 2007 – 2010 in the original data file from five populations.

From the original house sparrow data set and the large Chinook salmon data set, we choose three of the largest populations, and a fourth population which is defined as an outsider population. First, we compute the SNP genotype frequencies of these populations and set the values of the genotype matrix F to these values. Note that now $K = 4$ for both data sets but the number of SNPs J is 1322 for the empirical house sparrow data and 1242 for the empirical Chinook salmon data (only loci with non-missing entries are included).

After model parameters have been determined and fixed, we simulate SNP data following the scenario described in Anderson et al. (2008) (see their article Fig. 1 (a)). We simulate SNP data for four large populations with 500 individuals in each of the populations following the model (3). Finally, we randomly sample 100 baseline individuals from three populations of the size of 500 individuals. We also sample $m = 10$, 20 and 50 mixture individuals from the same three populations. In addition, we randomly sample 10 mixture individuals from the fourth "outsider" population: these mixture individuals do not have any genetic relationship with those 300 baseline individuals sampled from the other three simulated populations. Thus there are individuals from four populations in the mixture data but baseline data has only individuals from three populations. See supplementary Fig "Sampling" for schematic illustration of our mixture and baseline sampling procedure.

2.5.2 Data partitioning scheme

We use the procedures of the `RUBIAS` package to sample test data for a cross-validation style method comparison.

We apply all competing methods to sampled data sets that we generate using small subsets of individuals in:

1. The same house sparrow data we use in the data simulation scheme. Briefly, eight (8) different populations correspond to different islands on the coast of Norway, and the data used here was collected in year 2012. Overall, there are 183,145 SNPs and 507

baseline birds for 2012 in the original data file. The eight populations in the baseline data set had sample sizes 16, 20, 58, 63, 147, 130, 53 and 20.

2. Six (6) populations from a small Chinook salmon baseline data with under one hundred SNPs (Clemento et al., 2014). This data is freely available in the `RUBIAS` package and in Dryad digital repository (`https://doi.org/10.5061/dryad.574sv`). Overall there are 91 SNPs and 909 baseline fish for 2010 in the data file. The six populations had sample sizes 119, 95, 146, 117, 295 and 137. Genetic data was generated on a Fluidigm EP1 platform using 96.96 Dynamic Arrays.

We use the baseline population proportions from these data sets as the parameters of Dirichlet distribution and we treat the parameter vector simulated from the Dirichlet distribution as the "true" mixture proportions. Mixture individuals are randomly sampled from the baseline to satisfy these mixture proportions. Baseline samples are divided into a baseline group (eight populations in the house sparrow data and six populations in the Chinook salmon data) and a mixture group now with known origin. In this comparative analysis, we inspect how the cross-validation error of probability of the origins and mixture proportions differ when they are estimated with BONE, ADMIXTURE, RUBIAS and AssignPOP.

We sample $m = 100$ mixture individuals from both the sparrow and salmon baseline groups. We keep the rest of the individuals as baseline samples. The baseline and mixture sample sizes may vary slightly due to the random sampling done with the RUBIAS procedures. For computational convenience and to reduce dependence among SNPs of the house sparrow data, we selected 1000 SNPs from the sparrow data by systematically taking every 183th SNP from the data set (with random starting position).

2.5.3 Method comparison

We compare BONE with ADMIXTURE software (Alexander et al., 2009) which is a widely used software in population structure analysis. In particular, we exploit known ancestral populations in a supervised learning model in ADMIXTURE (Alexander and Lange, 2011) and use the EM algorithm for estimating the ancestry coefficient matrix Q. In addition, we compare our method with two recent population assignment techniques, RUBIAS (Moran

and Anderson, 2019) and AssignPOP (Chen et al., 2018). These methods are implemented in R packages `radmixture`, `RUBIAS` and `AssignPOP` respectively. With RUBIAS, we set allele frequency priors to constant scales and ran 1000 MCMC iterations with a "burn-in" set to 100. Then we computed posterior mean estimates based on all remaining MCMC iterations. With assignPOP, we used two classifiers: support vector machines (SVM) and naive Bayes. With BONE, we pre-set the LASSO tuning parameter to a decreasing sequence from 0.4 to 0.02 of the length 40 (on a non-log scale). Because BONE depends on the R `glmnet` package (Friedman et al., 2010) that cannot handle missing data, we removed markers with more than 5% missing values and imputed the rest of the missing genotype values with marker mode for BONE. This was repeated in each sampling round. On average, we had to remove 0.7% percent of the SNPs and 68.0% contained at least one missing genotype that had to be imputed in the house sparrow data. In the small Chinook salmon data set of Clemento et al. (2014) the corresponding proportions were 1.4% and 13.6% respectively. For the results using alternative imputation strategy in BONE, see supplementary materials.

We compute the MSE for both probability of the origin and mixture proportion estimates over 50 simulation replications of both empirical and simulated data. The MSE for any matrix M is

$$MSE(\widehat{M}, M) = \frac{1}{50}\sum_{i=1}^{50}(\widehat{M} - M)^2. \tag{4}$$

When interpreting simulation analyses results, outgroup individuals were omitted while computing the MSE of both the probability of the origin and mixture proportions.

For comparison, we computed estimates for the probability of the origin and mixture proportions randomly: probability of the origin for each individual are sampled from Dirichlet distribution $Dir(\alpha)$ where the vector α corresponds to the baseline population proportions. Mixture proportions are also simulated from the $Dir(\alpha)$ distribution.

We computed MSEs for all the methods and summarized their values in Fig 3 and for the simulated data sets and Fig 6 for the partitioned empirical data sets. We note that due to possible dispersers in natural populations (see e.g., Pärn et al., 2012), the MSE of estimated probability of the origin, mixture proportions of the partitioning scheme cannot

be exactly zero: Assignments may be biased by complex genetic histories which make baseline populations less genetically divergent.

Individuals are assigned to baseline populations based on the largest estimate of probability of the origin (see also supplementary materials). We also tested how increasing the number of SNPs simulated from the Dirichlet-multinomial model changes the assignment accuracy and the estimates of the mixture proportions of solution path and "Winner Takes it All" methods.

To illustrate how BONE is able to identify outsiders from the simulated data sets, we select 10 of the most obvious outsiders (top 10 lowest MSE between the estimated and the expected probability of the origin) detected by BONE. We note that out of 90 simulated mixture individuals, 10 are from "outsider" population in our simulation scheme. In our example, true positive (TP) is a mixture individual sampled from outsider population and identified as an outsider. False positive (FP) is a mixture individual sampled from other three populations (with sample sizes 10, 20 and 50) and identified as an outsider. Thus the total number of non-outsiders is 80. We compute the averaged true positive rate (TPR), TP/10 and false positive rate (FPR), FP/80 over 50 simulation replications. A graphical representation of the estimated probability of the origins of mixture individuals (including 10 outsiders) in one random run of our qualitative outsider detection procedure is represented in Fig 4.

3 Results

As shown in Fig 3, all methods estimate the probability of the origin and the mixture proportions of simulated mixture individuals far better than random guessing. Fig 4 is graphical representation of the estimated probability of the origins of the simulated data (large Chinook salmon data as the starting point) with a randomly sampled data set. ADMIXTURE produces estimates of the probability of the origin and mixture proportions with the lowest averaged MSE values for simulated data. This is not surprising considering binomial proportions (3) which were used to simulate SNP genotype data come from the ADMIXTURE model. Nevertheless, averaged accuracy rates (proportion of simulated mixture individuals assigned to the correct baseline population) of the assignments of differ-

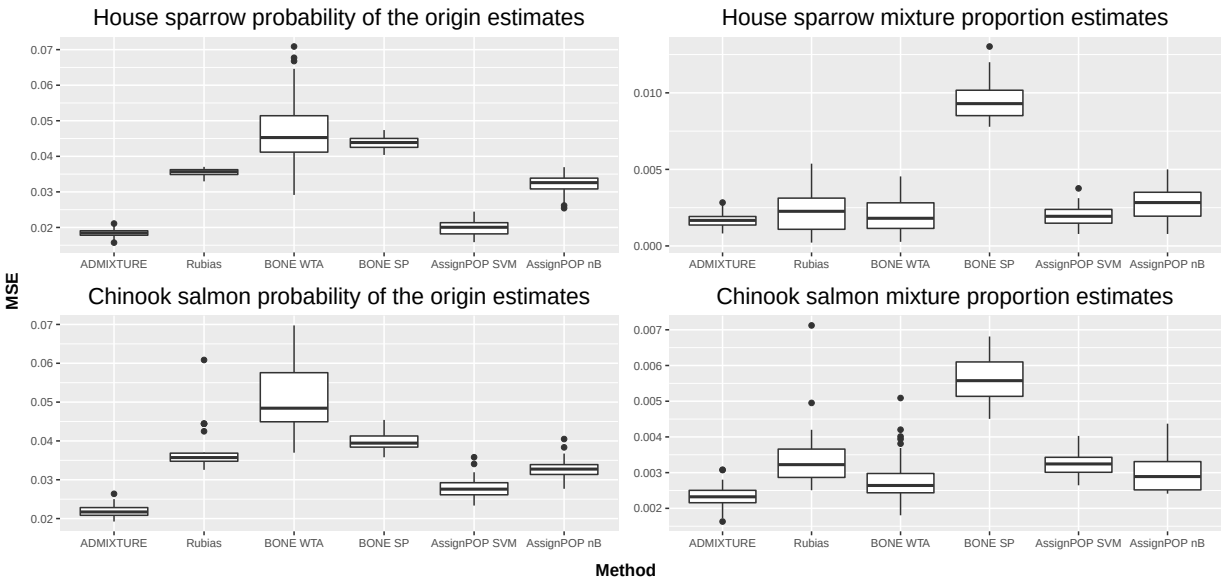

Figure 3: The mean squared errors (MSE) for different methods in the probability of the origin and the mixture proportions recovery based on 50 simulation replicates (using house sparrow and large Chinook salmon data SNP genotype frequencies as starting point). Here nB stands for naive Bayes, SVM for support vector machines, WTA for Winner Takes it All and SP for solution path. When estimates were determined from random samples of the Dirichlet distribution using true model parameters (random guessing), MSE median (standard deviation) of the probability of the origin and *mixture proportion* were 0.339 (0.049) and *0.061 (0.089)* along with 0.331 (0.039) and *0.067 (0.093)* for the house sparrow and Chinook salmon simulations respectively.

ent methods were very high: (i) simulated sparrow data: 100 per cent for BONE (solution path), ADMIXTURE and RUBIAS and 99.97 for AssignPOP (both naive Bayes and SVM) and 97.05 for BONE (WTA) (ii) simulated large Chinook data: 100 per cent for ADMIXTURE, 99.9 for BONE (solution path) and AssignPOP (both naive Bayes and SVM), 99.8 for RUBIAS and 96.03 for BONE (WTA).

The difference between averaged MSE values (Fig 3) can be partially explained by the number of available SNPs. To test this, we dropped out mixture individuals from the outside population and ran BONE analysis multiple times with different numbers of simulated markers. When the number of SNPs increases, the assignment accuracy improves and the MSE of the mixture proportions decreases in both BONE solution path and WTA

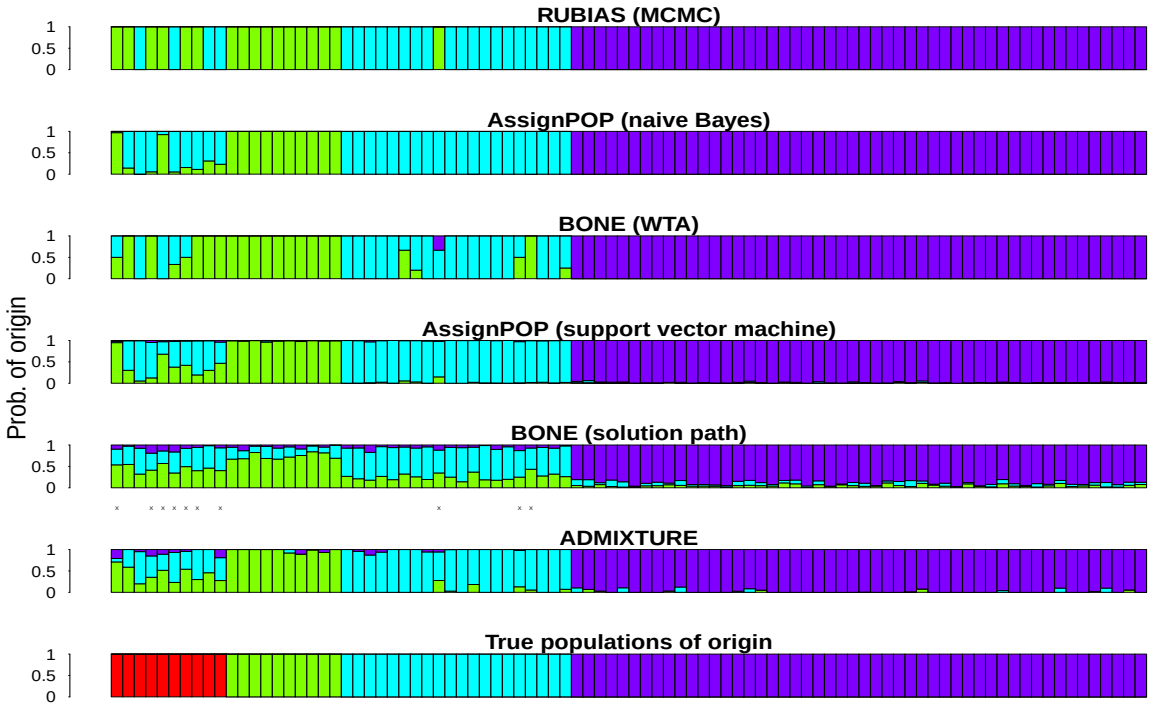

Figure 4: Graphical illustration of the probability of the origin for the 90 mixture individuals in one random data set among the 50 simulated data (large Chinook salmon data SNP genotype frequencies as starting point). "WTA" is short for "Winner Takes it All". Different colors correspond to different simulated baseline populations. The red population corresponds to so called outsider population with no correspondence in the baseline. Potential outsiders (10 individuals) are marked with an "X" in the network solution path bar chart.

methods (Fig 5). Using at least ca. 20,000 markers in the simulation analysis reduces the bias of the mixture proportion estimates practically to zero and the assignments are flawless.

When we specifically examine for the ability of BONE solution path method to detect outsiders in the simulated data sets, the averaged TPR and FPR are (i) 89.8 (standard deviation 6.5) and 1.3 (0.1) for the house sparrow simulated data, and (ii) 73.6 (standard deviation 11.2) and 3.3 (1.4) for the large Chinook simulated data, respectively. These 10 "outsider" mixture individuals are the ones among the mixture individuals having the

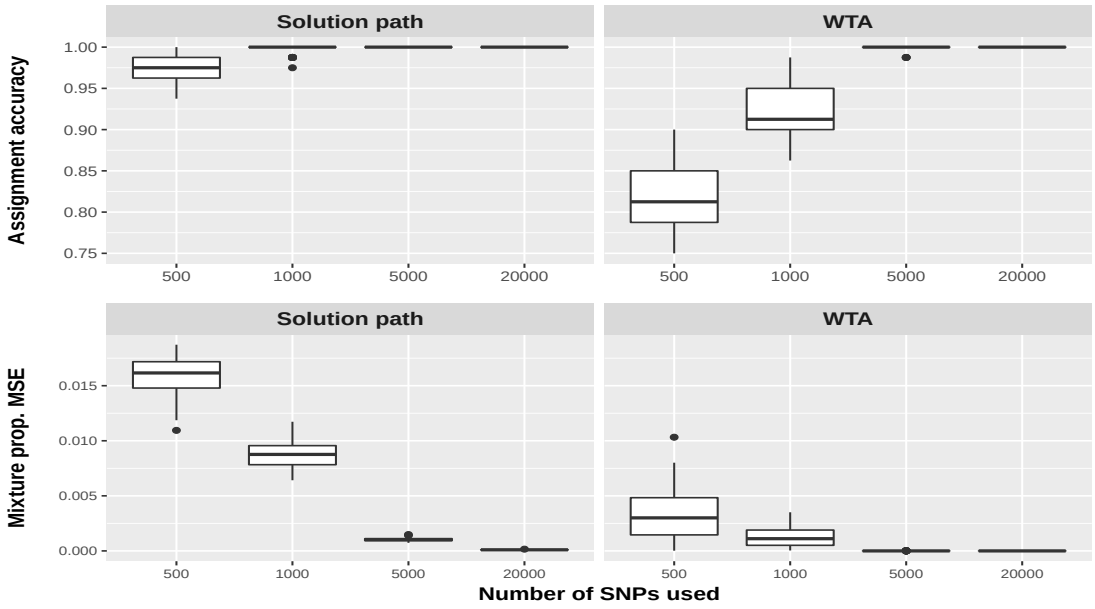

Figure 5: The assignment accuracy and the mean squared errors (MSE) of mixture proportions recovery based on 50 simulation replications (house sparrow data SNP genotype frequencies as the starting point) show the effect of number of SNP markers in the BONE method ("WTA" is short for "Winner Takes it All"). Assignment accuracy represents the proportion of correctly assigned mixture individuals.

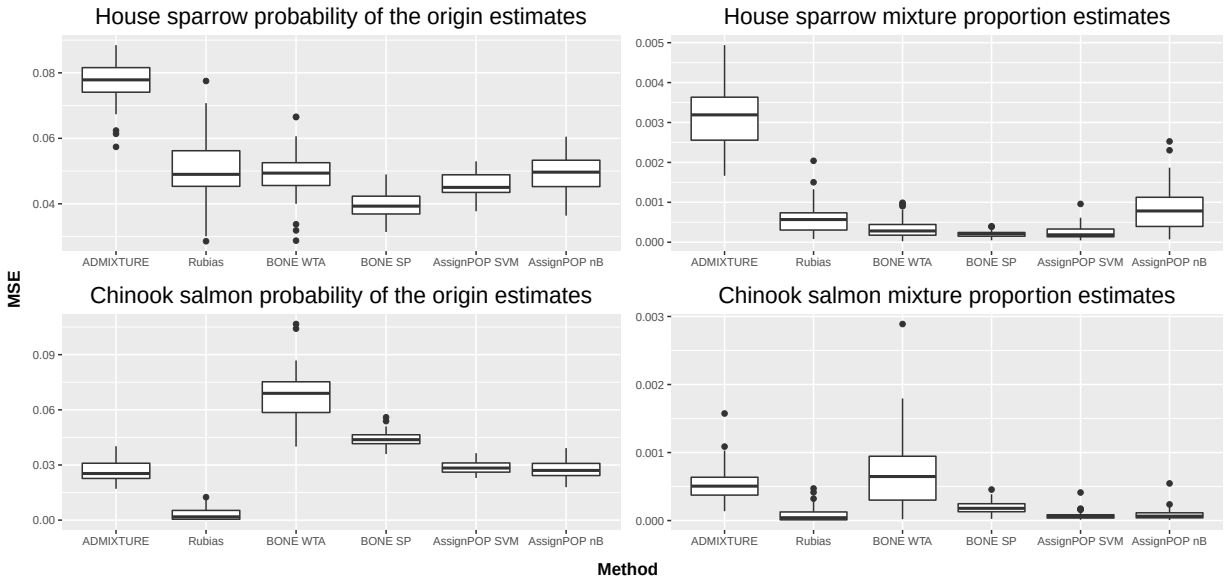

Figure 6: The mean squared errors (MSE) for different methods in the probability of the origin and the mixture proportions recovery based on 50 data partitioning replicates (empirical house sparrow data, 1,000 SNPs and small empirical Chinook salmon data, 91 SNPs). Here nB stands for naive Bayes, SVM for support vector machines, WTA for Winner Takes it All and SP for solution path. For comparison, when estimates were determined from random samples of the Dirichlet distribution using true model parameters (random guessing), MSE median (standard deviation) of the probability of the origin and *mixture proportion* were 0.118 (0.000) and *0.042 (0.031)* along with 0.148 (0.000) and *0.066 (0.038)* for the house sparrow and Chinook salmon simulations respectively.

weakest genetic similarity with baseline individuals compared to other individuals. LASSO model selection algorithm suppresses the genetic signals (covariate coefficients of the multinomial logit model, which define the neighbors in the baseline group of these mixture individuals, are shrunk exactly to zero by the LASSO penalty) of these individuals. Nevertheless, these results should be studied in more detail, because there is no guarantee that populations in the baseline group are genetically homogeneous, like in this simple example, when one examines true empirical data.

In Fig 6, when empirical sparrow and salmon data sets are examined, BONE produces the lowest MSE estimates with the house sparrow data (1,000 SNPs) and RUBIAS with the small Chinook salmon data set (91 SNPs). BONE is not accurate when the number of SNPs is very low although the averaged MSE values in Fig 6 indicates that mixture proportions of the small Chinook data are comparable with those produced by ADMIXTURE. Probability of the origin estimates of the house sparrow data computed with ADMIXTURE are fragmented and hard to interpret. This might be reflecting the presence of close relatives in the house sparrow data set or uneven sample sizes (Puechmaille, 2016; Wang, 2017; Lawson et al., 2018). See supplementary materials for a graphical illustration of the probability of the origin estimates of the empirical house sparrow data ("Supp_SparrowBarplots").

Finally, we note that a more detailed comparison between BONE and GSI_SIM (Anderson et al., 2008), which is a Unix-variant of RUBIAS, is presented in Saatoglu et al. (2019), where the house sparrow data that was used as a starting point to compare different methods in the current study, is analysed in its entirety.

3.1 Qualitative comparison of population assignment methods applied in this article

All population assignment methods we have examined here have their pros and cons. Most of their unwanted features relate to the implementation of these methods, and could be solved if their implementation involved more efficient programming or by using top-notch hardware. For example, the supervised learning model of ADMIXTURE implemented in the R package `radmixture` supports only one unknown individual at a time. These technical problems may be overcome via the evolution of computer hardware. We have collected

a small summary of the properties of RUBIAS, BONE, AssignPOP and ADMIXTURE methods in Table 1.

We note that unbalanced population sample sizes are often the case in empirical studies (like in our house sparrow data set), and they are problematic for some methods (see, e.g., Puechmaille, 2016; Wang, 2017). Our previous study with the multinomial logit model suggests that our graph method is not sensitive to uneven sample sizes (Kuismin et al., 2017).

	RUBIAS	BONE	AssignPOP	ADMIXTURE
Running time	Fast	Moderate	Fast	Fast (slow[a])
Initial costs[b]	Moderate	Moderate	High	Small
Performance with small data sets	Good	Moderate	Good[c]	Good
Performance with large data sets	Good	Good	Good	Good
Outsider detection	No	Yes	No	No
Effect of unbalanced sample sizes	?	Negligible[d]	Negligible	High[e]

[a] The original command line program is fast but the R implementation `radmixture` only supports estimation of one individual with unknown ancestry at a time.

[b] Time needed for reading and formatting the raw genotype data.

[c] AssignPOP provides tools to reduce the number of loci with low variance.

[d] Kuismin et al. (2017).

[e] Puechmaille (2016); Wang (2017).

Table 1: A qualitative comparison of the properties of RUBIAS, BONE, AssignPOP and ADMIXTURE.

In addition to the features summarized in Table 1, there are other practical differences. For example, each method uses different file formats: AssignPOP uses either GENEPOP formatted files (Rousset, 2008) or STRUCTURE formatted files (Pritchard et al., 2000), RUBIAS uses its own file formats, and BONE was initially developed for PED formatted files used by, e.g. PLINK (Purcell et al., 2007). Previously published methods have not been able to detect outsiders, but it may be possible to overcome this limitation in them by using our suggested detection procedure based on the MSEs of the probability of the

origin estimates.

4 Conclusions

Here we have provided a detailed description of how to apply graph estimation tools in genetic assignment. We have shown that BONE produces probability of the origin and mixture proportion estimates, which are comparable or superior to recently published genetic assignment methods. Furthermore, general characteristics of the network model (i.e., nodes might have no neighbors) provide a separate group for those individuals which do not seem to have any logical assignment to baseline populations (i.e., no mixture individual is forced into a baseline population). Therefore, we have shown how these outsiders can be correctly identified. A characteristic of BONE is that one can actually inspect the baseline neighbors of each mixture individual from a graph.

There is still room for improvement of our graph method. For example, the genetic relationship between mixture individuals and baseline individuals is defined by binary values in the adjacency matrix. The smaller variance associated with the estimates computed with our solution path method in MSE sense seem to indicate that a weighted graph could describe the genetic relationship between mixture and baseline individuals in more detail. Moreover, if this relationship could be defined with weighted values, $A = [a_{i,j}]$, $0 \leq a_{i,j} \leq 1$, it is theoretically possible to determine the probability of the origin and mixture proportions with minimal estimation error. However, this is a far more challenging graphical model to estimate and we leave its development and examination for future studies.

BONE avoids the difficult task of LASSO tuning parameter selection, which makes BONE practically parameter free. The user does not have to be an expert of the LASSO-regression to use BONE. The only serious limitation, where BONE estimates are biased, is when an individual has very low heterozygosity or homozygosity. Compared to ADMIXTURE, RUBIAS and AssignPOP, BONE also allows inspection of which baseline individuals are the most probable close relatives of a given mixture individual by inspecting the nodes of the network (see the supplementary Fig "WTANetworkEstimate"). We leave a thorough investigation of this property for future studies.

We have provided example data sets and open source code for users to apply BONE in

practice. BONE provides a methodological starting point and a framework which is a new addition and useful alternative to toolbox of existing genetic assignment techniques.

Acknowledgments

This work was supported by the University of Oulu's Exactus Doctoral Programme, Biocenter Oulu funding, the Technology Industries of Finland Centennial Foundation, the Jane and Aatos Erkko Foundation, and by grants from the Research Council of Norway (projects 221956 and 274930) and the Academy of Finland (project 295204 to A.N.). This work was also partly supported by the Research Council of Norway through its Centres of Excellence funding scheme (project 223257). The house sparrow fieldwork was carried out in accordance with permits from the Ringing Centre at Stavanger Museum, Norway. SNP genotyping on the custom house sparrow Affymetrix Axiom 200K SNP array was carried out at CIGENE, Norwegian University of Life Sciences, Norway. We have no conflict of interest to declare.

Author contributions

All authors were involved in the conception and design of the method. M.K. and M.J.S. developed the software. M.K. executed the analyses. D.S. A.N. and H.J. provided the empirical house sparrow data. All authors interpreted results and critically revised the manuscript.

Data Availability

The BONE R package for genetic assignment of individuals, a demo script, an example data set and collection of scripts used to prepare the material in this paper are publicly available at GitHub under the GPL license (https://github.com/markkukuismin/BONE) and deposited at Zenodo (https://doi.org/10.5281/zenodo.3517785). Phenotypic data, alongside SNP genotype data for house sparrows is available in the Dryad repository: https://doi.org/10.5061/dryad.gqnk98sh8. The data have been anonymized at the request

of the data owners. Supporting Figures and detailed descriptions of methodology are uploaded as a supplementary files in Supporting Information.

ORCID

Markku Kuismin https://orcid.org/0000-0001-9074-7420
Dilan Saatoglu https://orcid.org/0000-0003-3936-287X
Alina K. Niskanen https://orcid.org/0000-0003-2017-2718
Henrik Jensen https://orcid.org/0000-0001-7804-1564
Mikko J. Sillanpää https://orcid.org/0000-0003-2808-2768

References

Alexander, D. H. and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12:246.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664.

Anderson, E. (2010). Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, 10:701–710.

Anderson, E. C., Waples, R. S., and Kalinowski, S. T. (2008). An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, 65:1475–1486.

Araya-Ajoy, Y. G., Ranke, P. S., Kvalnes, T., Rønning, B., Holand, H., Myhre, A. M., Pärn, H., Jensen, H., Ringsby, T. H., Sæther, B.-E., and Wright, J. (2019). Characterizing morphological (co)variation using structural equation models: Body size, allometric relationships and evolvability in a house sparrow metapopulation. *Evolution*, 73:452 – 466.

Beacham, T. D., Jonsen, K., and Wallace, C. (2012). A comparison of stock and individual identification for Chinook salmon in British Columbia provided by microsatellites and single-nucleotide polymorphisms. *Marine and Coastal Fisheries*, 4:1–22.

Chen, K.-Y., Marschall, E. A., Sovic, M. G., Fries, A. C., Gibbs, H. L., and Ludsin, S. A. (2018). assignPOP: An R package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework. *Methods in Ecology and Evolution*, 9:439–446.

Clemento, A. J., Crandall, E. D., Garza, J. C., and Anderson, E. C. (2014). Evaluation of a single nucleotide polymorphism baseline for genetic stock identification of Chinook Salmon (Oncorhynchus tshawytscha) in the California Current large marine ecosystem. *Fishery Bulletin*, 112(2-3):112–130.

Clobert, J., Baguette, M., Benton, T. G., and Bullock, J. M. (2012). *Dispersal Ecology and Evolution*. Oxford University Press, Oxford, UK.

Driscoll, D. A., Banks, S. C., Barton, P. S., Ikin, K., Lentini, P., Lindenmayer, D. B., Smith, A. L., Berry, L. E., Burns, E. L., Edworthy, A., Evans, M. J., Gibson, R., Heinsohn, R., Howland, B., Kay, G., Munro, N., Scheele, B. C., Stirnemann, I., Stojanovic, D., Sweaney, N., Villaseñor, N. R., and Westgate, M. J. (2014). The trajectory of dispersal research in conservation biology. Systematic review. *PLOS ONE*, 9(4):1–18.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393.

Dyer, R. J. and Nason, J. D. (2004). Population graphs: the graph theoretic shape of genetic structure. *Molecular Ecology*, 13:1713–1727.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

Garroway, C. J., Bowman, J., Carr, D., and Wilson, P. J. (2008). Applications of graph theory to landscape genetics. *Evolutionary Applications*, 1:620–630.

Garvin, M. R., Saitoh, K., and Gharrett, A. J. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, 10:915–934.

Greenbaum, G., Templeton, A. R., and Bar-David, S. (2016). Inference and analysis of population structure using genetic data and network theory. *Genetics*, 202:1299–1312.

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M., Ogden, R., Limborg, M., Cariani, A., Maes, G., Diopere, E., Carvalho, G., et al. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, 11:123–136.

Horvath, S. (2011). *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer Science & Business Media, New York, USA.

Huson, D. H. and Scornavacca, C. (2011). A survey of combinatorial methods for phylogenetic networks. *Genome Biology and Evolution*, 3:23–35.

Jiang, Z., Wang, H., Michal, J. J., Zhou, X., Liu, B., Woods, L. C. S., and Fuchs, R. A. (2016). Genome wide sampling sequencing for SNP genotyping: methods, challenges and future development. *International Journal of Biological Sciences*, 12:100–108.

Kuismin, M. O., Ahlinder, J., and Sillanpää, M. J. (2017). CONE: Community oriented network estimation is a versatile framework for inferring population structure in large scale sequencing data. *G3 (Bethesda)*, 7:3359–3377.

Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D., and Seeb, J. E. (2014). Genotyping by sequencing resolves shallow population structure to inform conservation of chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, 7:355–369.

Lawson, D. J., Van Dorp, L., and Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9:3258.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., et al. (2008). Worldwide

human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, 42:413–468.

Lundregan, S. L., Hagen, I. J., Gohli, J., Niskanen, A. K., Kemppainen, P., Ringsby, T. H., Kvalnes, T., Pärn, H., Rønning, B., Holand, H., Ranke, P. S., Båtnes, A. S., Selvik, L.-K., Lien, S., Sæther, B.-E., Husby, A., and Jensen, H. (2018). Inferences of genetic architecture of bill morphology in house sparrow using a high-density SNP array point to a polygenic basis. *Molecular Ecology*, 27:3498–3514.

Manel, S., Gaggiotti, O. E., and Waples, R. S. (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, 20:136–142.

McKinney, G. J., Seeb, J. E., and Seeb, L. W. (2017). Managing mixed-stock fisheries: genotyping multi-SNP haplotypes increases power for genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, 74:429–434.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:417–473.

Moran, B. M. and Anderson, E. C. (2019). Bayesian inference from the conditional genetic stock identification model. *Canadian Journal of Fisheries and Aquatic Sciences*, 76:551 – 560.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., et al. (2008). Genes mirror geography within europe. *Nature*, 456:98–101.

Pärn, H., Ringsby, T. H., Jensen, H., and Sæther, B.-E. (2012). Spatial heterogeneity in the effects of climate and density-dependence on dispersal in a house sparrow metapopulation. *Proceedings of the Royal Society B: Biological Sciences*, 279:144–152.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Puechmaille, S. J. (2016). The program `STRUCTURE` does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, 16:608–627.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81:559–575.

Ronce, O. (2007). How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annual Review of Ecology, Evolution, and Systematics*, 38:231–253.

Rousset, F. (2008). GENEPOP'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, 8:103–106.

Ruegg, K. C., Anderson, E. C., Harrigan, R. J., Paxton, K. L., Kelly, J. F., Moore, F., and Smith, T. B. (2017). Genetic assignment with isotopes and habitat suitability (GAIAH), a migratory bird case study. *Methods in Ecology and Evolution*, 8:1241–1252.

Saastamoinen, M., Bocedi, G., Cote, J., Legrand, D., Guillaume, F., Wheat, C. W., Fronhofer, E. A., Garcia, C., Henry, R., Husby, A., Baguette, M., Bonte, D., Coulon, A., Kokko, H., Matthysen, E., Niitepõld, K., Nonaka, E., Stevens, V. M., Travis, J. M. J., Donohue, K., Bullock, J. M., and del Mar Delgado, M. (2018). Genetics of dispersal. *Biological Reviews*, 93:574–599.

Saatoglu, D., Niskanen, A. K., Kuismin, M., Ranke, P. S., Hagen, I. J., Araya-Ajoy, Y., Myhre, A. M., Holand, H., Kvalnes, T., Pärn, H., Sommerli, S. L., Rønning, B., Lien, S., Ringsby, T. H., Sæther, B.-E., Husby, A., Sillanpää, M. J., and Jensen, H. (2019).

Dispersal in a house sparrow metapopulation - identifying "cryptic" dispersers using genetic assignment. *(Manuscript)*, 0:0–0.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288.

Wang, J. (2017). The computer program `STRUCTURE` for assigning individuals to populations: easy to use but easier to misuse. *Molecular Ecology Resources*, 17:981–990.

Wang, Y. R. and Huang, H. (2014). Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology*, 362:53–61.