BIOLOGY
Methods & Protocols

Biology Methods and Protocols, 2021, 1–6

doi: 10.1093/biomethods/bpab017
Innovations

INNOVATIONS

OptM: estimating the optimal number of migration edges on population trees using *Treemix*

Robert R. Fitak *

Department of Biology, Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816, USA

*Correspondence address. Department of Biology, Genomics and Bioinformatics Cluster, University of Central Florida, 4110 Libra Dr., Orlando, FL 32816, USA. E-mail: Robert.fitak@ucf.edu

Abstract

The software *Treemix* has become extensively used to estimate the number of migration events, or edges (m), on population trees from genome-wide allele frequency data. However, the appropriate number of edges to include remains unclear. Here, I show that an optimal value of m can be inferred from the second-order rate of change in likelihood (Δm) across incremental values of m. Repurposed from its original use to estimate the number of population clusters in the software *Structure* (ΔK), I show using simulated populations that Δm performs equally as well as current recommendations for *Treemix*. A demonstration of an empirical dataset from domestic dogs indicates that this method may be preferable in large, complex population histories and can prioritize migration events for subsequent investigation. The method has been implemented in a freely available R package called "OptM" and as a web application (https://rfitak.shinyapps.io/OptM/) to interface directly with the output files of *Treemix*.

Keywords: likelihood; population genomics; SNPs; structure

Introduction

One of the fundamental aspects of modern population genetics is using allele-frequency measurements to recreate the various demographic events that define an extant species. However, species and their constituent populations often contain complex demographic histories that may include various instances and fluctuations in migration, population size, and fragmentation. These complex demographic scenarios often require large amounts of genetic data to be sufficiently resolved. Recent advances in sequencing and genotyping technologies [notably for single-nucleotide polymorphisms (SNPs)] have made the generation of genome-wide allele frequency data for multiple populations increasingly tractable [1], thus limiting studies of demographic history primarily to the statistical models and computational capabilities available.

A graph-based model for describing the relationships between populations was recently described by Pickrell and Pritchard [2]. This approach is able to estimate population splits and migration by first building a tree model of the populations then subsequently adding migration events (or edges) between populations that poorly fit the tree model. Pickrell and Pritchard implemented their model in a software package called *Treemix*, which has been used to infer gene flow between populations of many species including fungi (e.g., [3]), plants (e.g., [4]), reptiles (e.g., [5]), mammals (e.g., [6–9]), and numerous others. *Treemix* allows the user to model any number of migration edges, and the authors suggested that, based upon simulations, a model that explains 99.8% of variation in the relatedness between populations is sufficiently robust to infer the number of migration edges. Nevertheless, real-world demographic histories are often more complex and this approach may underestimate or overestimate the number of migrations edges. For example, when the ratio of the number of admixed to unadmixed populations is quite large, *Treemix* may simply account for this by shortening the branch to the unadmixed populations in the tree rather than adding multiple migration edges [2].

Received: 20 August 2021; **Revised:** 9 September 2021; **Editorial Decision:** 10 September 2021; **Accepted:** 13 September 2021

In this study, I propose the application of an *ad hoc* statistic similar to the method described by Evanno *et al.* [10] for the software *Structure* [11] to determine the optimal number of migration events to include when using *Treemix*. I demonstrate on simulated populations that this approach performs equally as well as the 99.8% variation threshold suggested by Pickrell and Pritchard [2]. However, using an empirical example of domestic dogs and wolves, I show the utility of this approach under large, complex demographic histories when the recommended threshold becomes difficult to obtain.

Materials and methods

Approach

Inferring the most probable number of migration events in a model, or m, is akin to inferring the most likely number of populations, K, when using the software *Structure* [11]. *Structure* is perhaps the most widely used program for inferring population structure and admixture [12], and over the past two decades has become a standardized tool for population genetic studies [13]. When using *Structure*, it was recommended to infer the most likely value for K by (i) performing multiple runs with various values for K, (ii) plotting the log posterior probability of the data given K ("$\ln P(D)$," or simply "$L(K)$") for each run, and (iii) observing the value of K where $L(K)$ reaches a plateau and/or the variance begins to increase [14]. However, Evanno *et al.* [10] demonstrated that this approach may not be accurate and proposed a new method that improves the ability to predict the true value of K. The method proposed by Evanno *et al.* calculated an *ad hoc* statistic, ΔK, based upon the second-order rate of change in $L(K)$.

I propose here that the same procedure proposed by Evanno *et al.* can be used to estimate the most likely value for m when using *Treemix*. The only difference is m can have values ≥ 0 whereas K must be ≥ 1. The software *Treemix* calculates composite log-likelihoods for each run using models both without migration edges ($m = 0$) and with m edges. I define these likelihoods as $L(m)$ and they are analogous to the $L(K)$ values produced by *Structure*. By performing multiple runs with different values for m, the same methodology to calculate ΔK can be used to calculate its migration equivalent, Δm. I refer users to Evanno *et al.* [10] and Supplementary File S1 for a complete description of the model and its calculations. I have implemented the method in a software package called *OptM* v0.1.5 available for the R programming language through CRAN [15] (https://cran.r-project.org); designed specifically for use with the output files produced by *Treemix* v1.13 (https://bitbucket.org/nygcresearch/treemix; RRID: SCR_021636). *OptM* was built originally using R v3.2.2, but has been tested extensively to function properly on various platforms (i.e., Windows and Unix) through the current version R v4.1.1. Its dependencies include the packages SiZer ($\geq$v0.1-4), stats, splines, grDevices, and boot ($\geq$v1.3-20). With regard to the input files, *OptM* analyzes all the *Treemix* output files in a given folder with the suffixes ".llik," ".cov.gz," and ".modelcov.gz" generated by default using *Treemix*. *OptM* generates an output table with the calculations and an estimated optimal value of m and includes a plotting function "*plot_optM*" to visualize the results and produce publication-ready figures. Alternatively, *OptM* incorporates added functionality to estimate m using change points estimated from threshold models often employed in ecology (see Ref. [16] and Supplementary File S1). *OptM* can fit parametric models such as piecewise linear, bent cable, simple exponential, and non-linear least squares to the $L(m)$ values across runs and compare them with the Akaike information criterion [17]. The non-parametric "significant zero crossings" method (SiZer) [16] is also available for comparison purposes.

Testing using simulations

I generated four simulated datasets each comprising 20 populations that evolved according to a serial bottleneck scenario using the whole-genome coalescent simulator *Argon* v0.1 (https://palamaralab.github.io/software/argon/; RRID: SCR_021635) [18]. The population graph without any migration events was identical to that in Pickrell and Pritchard [2] (Supplementary Fig. S1). The simulations each assumed an effective population size of 10^4 and all populations are descended from a common ancestor 2000 generations in the past. However, within each simulated dataset, 1, 3, 5, or 8 migration events were included 100 generations in the past. The source and recipient population for each migration event were selected at random without replacement, and the recipient population received 30% of its genetic ancestry from the source population. Each simulation produced 60 chromosomes (30 diploid individuals) of 250 megabases for each of the 20 populations. The simulations included a mutation rate of 10^{-8} substitutions $\cdot$ site^{-1} $\cdot$ generation^{-1} and recombination rate of 10^{-8} recombination $\cdot$ site^{-1} $\cdot$ generation^{-1}. The simulation parameters were identical to that of Pickrell and Pritchard [2]—resulting in patterns of diversity and linkage disequilibrium consistent with that of SNP genotype data for many modern human datasets [19] (see Supplementary File S1). To further recapitulate patterns in observed datasets, all loci with a

Table 1: Admixture proportions inferred by *Treemix* for 1 (M1), 3 (M3), 5 (M5), or 8 (M8) simulated migration edges (rows)

Migration edge	M1	M3	M5	M8
13 → 5	0.30 (0.0011)	0.30 (0.00062)	0.32 (0.0016)	0.29 (0.0036)
4 → 12		0.29 (0.00015)	0.27 (0.0014)	0.28 (0.00024)
7 → 16		0.30 (0.00037)	0.31 (0.023)	0.31 (0.00024)
6 → 3			0.33 (0.0011)[a]	0.31 (0.00098)[b]
9 → 17			0.28 (0.0086)	0.29 (0.00027)
15 → 1				0.29 (0.076)
14 → 8				0.29 (0.00029)
19 → 10				0.30 (0.012)[c]

The direction of the simulated migration is reported in the first column (source → sink). The standard deviation from 10 iterations is shown within parentheses.

[a] *Treemix* incorrectly inferred the migration edge 11 → 4 rather than 6 → 3 for all 10 iterations.

[b] *Treemix* correctly inferred the 6 → 3 migration edge in 9/10 iterations, but one iteration incorrectly reported a 15 → 10 edge.

[c] In 9/10 iterations, *Treemix* reported the source of this migration edge to be the common ancestor of 19 and 20, or 19/20 → 10.

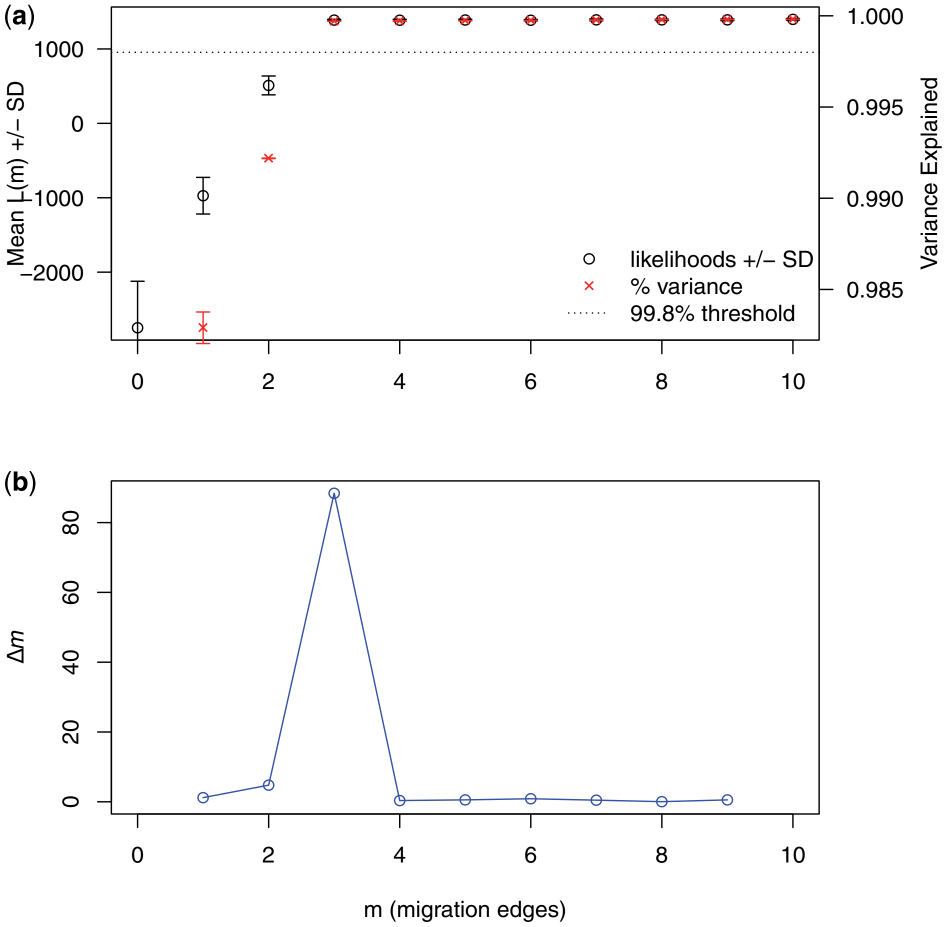


Figure 1: The output produced by *OptM* for the simulated dataset with $m = 3$ migration edges. (**a**) The mean and standard deviation (SD) across 10 iterations for the composite likelihood $L(m)$ (left axis, black circles) and proportion of variance explained (right axis, red "x"s). The 99.8% threshold (horizontal dotted line) is that recommended by Pickrell and Pritchard [2]. (**b**) The second-order rate of change (Δm) across values of m.

minimum allele frequency <0.05 were removed using *Vcftools* v0.1.13 (https://vcftools.github.io/index.html; RRID: SCR_001235) [20]. The resulting datasets were run using *Treemix* v1.13 with a global set of rearrangements (-global), and a randomly selected window size (-k) of between 100 and 1000 SNPs (50 SNP increments). The number of migration events (-m) varied between 1 and 10 [*Treemix* natively calculates the $L(m = 0)$ for each run] and 10 replicates were performed for each value of m. Although the parameter "-k" can be fixed across runs, the simulated datasets generally created strong signals that resulted in convergence upon the same composite likelihood across all runs for each value of m. This produced a standard deviation across runs of zero, and thus Δm becomes undefined. In this scenario *OptM* will generate a warning to the user, but the practice of permuting across "-k" or across the set of input SNPs will improve the reliability of estimates of Δm. The resulting likelihood files produced by *Treemix* were analyzed and visualized using the functions in *OptM*. See Supplementary File S1 for a complete description of the methods including computer code used.

Empirical example

I applied this method to an empirical dataset composed of 532 domestic dogs from 48 breeds and 15 wolves genotyped for ~174,000 SNPs on the CanineHD BeadChip [21, 22]. In order to accurately estimate allele frequencies, we removed breeds with less than eight individuals genotyped. The SNPs were filtered to include only autosomal loci with a minimum allele frequency ≥0.05 and a genotyping rate ≥0.9 using *Plink* v1.07 (https://zzz.bwh.harvard.edu/plink/; RRID: SCR_001757) [23]. Individuals with a genotyping rate ≤0.9 were omitted from the analysis. The resulting dataset was run using *Treemix* v1.13 with the same parameters as above with the exceptions of a window size (-k) of 500 SNPs and number of migration events (-m) between 1 and 40. Again, 10 replicates were performed for each value of m and the resulting files were analyzed using *OptM* (see Supplementary File S1).

Results and discussion

Simulated examples

Four simulated datasets containing either 1 (M1), 3 (M3), 5 (M5), or 8 (M8) migration events from a serial bottleneck model of 20 populations were generated (Supplementary Fig. S1). Each migration edge was simulated with 30% admixture (referred to as "migration weight," or $\hat{w}$, by Pickrell and Pritchard [2]), which *Treemix* was able to accurately infer (range of $\hat{w}$ 27% - 33%; Table 1) in all datasets. However, in M5, *Treemix* incorrectly reported a migration edge from populations 11 → 4 rather than populations 6 → 3 in all iterations, and in M8 a single iteration incorrectly assigned a migration edge between populations 15 → 10

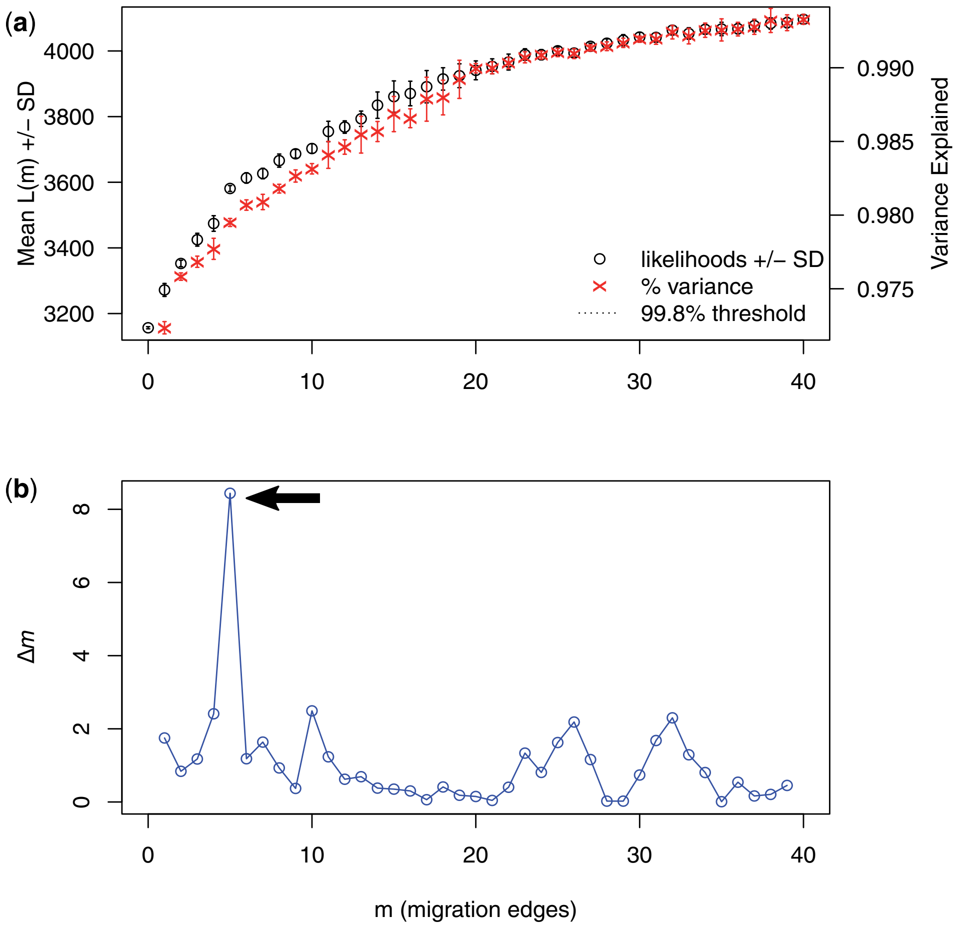


Figure 2: The output produced by *OptM* for an empirical dataset of domestic dogs. A total of 10 iterations were run for each possible number of migration edges, $m = 1$–40. (**a**) The mean and standard deviation (SD) for the composite likelihood $L(m)$ (left axis, black circles) and proportion of variance explained (right axis, red "x"s). The 99.8% threshold is that recommended by Pickrell and Pritchard [2], but not visible here because the threshold is still not met at $m = 40$ edges. *OptM* produces a warning to notify the user that this threshold is not visible. (**b**) The second-order rate of change (Δm) across values of m. The arrow indicates the peak in Δm at $m = 5$ edges.

Table 2: Summary of all the migrations edges inferred across 10 iterations at $m = 5$ (the optimal number inferred by *OptM*)

Source lineage	Recipient lineage	$\hat{w}$ (SD)	Number of iterations
Box	ShP	0.093 (0.0068)	9/10
Sci	Eur	0.41 (0.0036)	9/10
Sci	StP+PdL	0.41 (0.0048)	9/10
EBD	NSD+((LRe+GRe)+NFd)	0.085 (0.0023)	8/10
EBD+EBT	IrW+Gry	0.21 (0.0063)	7/10
EBT	IrW+Gry	0.16 (0.00094)	2/10
EBD+EBT	NSD+((GRe +LRe)+NFd)	0.11 (NA)	1/10
EBD+EBT	(Rtw+BMD)+(NSD+((GRe+LRe)+NFd))	0.086 (NA)	1/10
EBD	ShP	0.046 (NA)	1/10
EBD	IrW+Gry	0.15 (NA)	1/10
ShP	Box	0.10 (NA)	1/10
StP	(((GSl+(Wlf+(ShP+Eur)))+FSp)+Elk)+Sci	0.15 (NA)	1/10
StP	((((Wlf+(ShP+Eur))+GSl)+FSp)+Elk)+Sci	0.17 (NA)	1/10

$\hat{w}$, migration weight, Breed abbreviations are as follows: Box, Boxer; BMD, Burnese Mountain Dog; Elk, Elkhound; EBD, English Bulldog; EBT, English Bull Terrier; Eur, Eurasian; FSp, Finnish Spitz; GRe, Golden Retriever; GSl, Greenland Sledge Dog; Gry, Greyhound; IrW, Irish Wolfhound; LRe, Labrador Retriever; NFd, Newfoundland; NSD, Nova Scotia Duck Tolling Retriever; PdL, Poodle; Rtw, Rottweiler; Sci, Schipperke; ShP, Shar-Pei; StP, Standard Poodle; Wlf, wolf.

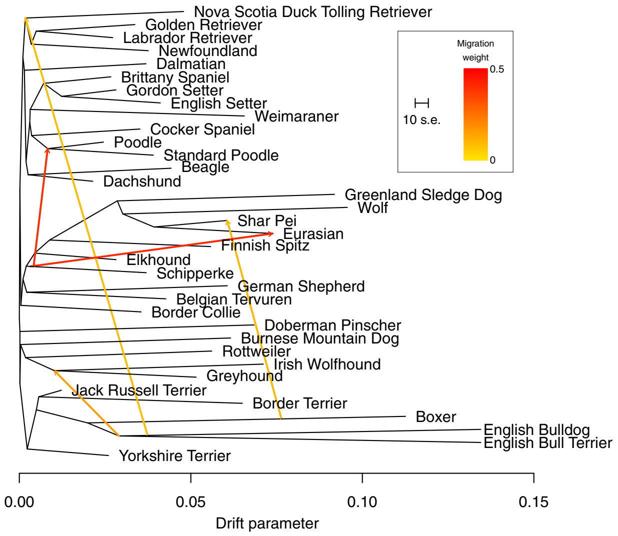


Figure 3: The tree structure of the graph inferred by *Treemix* for the 34 dog breeds and gray wolf populations. Five migration edges were allowed as inferred by *OptM*. The migration edges are colored according to their weight ($\hat{w}$). The scale bar indicates ten times the average standard error of the values in the covariance matrix.

rather than populations 6 → 3 (Table 1). In M8, *Treemix* reported in 9/10 iterations migration from the common ancestor of populations 19 and 20 into population 10, rather than populations 19 → 10 (Table 1). The latter error is understandable as populations 19 and 20 only separated ten generations previously. Nevertheless, *Treemix* was able to accurately and consistently infer the correct migration edges across the simulated datasets and is a testament to the utility of this algorithm in inferring population histories.

When using the method in *OptM* to infer the optimal value for m from 0 to 10 simulated migration edges, the highest value for the second-order rate of change in likelihood, or Δm, identified the correct number of simulated migration edges in all four datasets (Fig. 1, Supplementary Fig. S2). The inferences based on Δm for datasets M3, M5, and M8 were equivalent to those suggested by the authors of *Treemix* based upon the 99.8% variation cutoff [2]. However, for dataset M1, which had only one migration edge, the 99.8% threshold was exceeded even when no migration edges were inferred, but *OptM* was able to correctly identify this situation. As a result, *OptM* may outperform the previous method when the true number of migration edges is very small. *OptM* was also used to fit piecewise linear, bent cable, simple exponential, and non-linear least squares threshold models to $L(m)$ (Supplementary Table S1 and Fig. S3), albeit the Δm method outperformed these models.

Empirical example

I ran *Treemix* on an empirical dataset that, after filtering, contained 496 domestic dogs from 34 breeds, 12 wolves, and >138 000 SNPs. I ran 400 instances of *Treemix*, 10 iterations for $m = 1$–40 migration edges. Even after including 40 migration edges, the 99.8% recommended threshold for stopping the addition of migration edges was not met (Fig. 2a). Rather, *OptM* suggested that five migration edges should be optimally included (Figs 2b and 3).

Although each of the 10 iterations at $m = 5$ inferred a slightly different set of migration edges, five migrations edges were substantially more common than the others (Table 2). These included $\hat{w}$ = 9.3% (SD 0.68%) from the boxer into the Shar-Pei, similar to the $\hat{w}$ = 8% reported by Pickrell and Pritchard [2] for the same edge using a different set of SNPs. This edge is likely the result of the Shar-Pei being considered an ancient breed [24, 25] and the fact that most canine SNPs on commercial genotyping chips were ascertained from the boxer's genome [2, 22]. Extensive gene flow from the Schipperke into the Eurasian ($\hat{w}$ = 41% SD 0.36%) is consistent with the known European × East Asian spitz-type hybrid origin of the Eurasian and has been observed elsewhere [25]. Gene flow from the Schipperke into the Poodle ($\hat{w}$ = 41% SD 0.48%) is less clear, as this migration edge has yet to be described, but could be a result of the fact that 65.1% of SNPs on the CanineHD beadchip were ascertained from a Boxer-Poodle comparison [22]. The remaining two notable migration edges were from the English Bulldog into the ancestor of the Nova Scotia Duck Tolling Retriever, Labrador Retriever, Golden Retriever, and Newfoundland ($\hat{w}$ = 8.5% SD 0.23%) and both the English Bulldog + English Bull Terrier into the Irish Wolfhound and Greyhound ($\hat{w}$ = 21% SD 0.63%). All of the breeds included in these two migration edges originate from the British Isles, and probably illustrate the many cross-breeding events that took place to create hybrid varieties that would excel in dog fighting contests prior to the strict studbook keeping in the middle to late nineteenth century [8, 26, 27]. All migration edges inferred by *Treemix* appeared early on the various branches (Fig. 3), and thus most likely represent ancient gene flow that predates modern breed development and the complex nature of domestic dog evolutionary history.

Conclusions

I have demonstrated here that the method of Evanno *et al.* [10] developed for inferring the number of population clusters from *Structure* and implemented in *OptM* can be repurposed to infer the optimal number of migration edges using *Treemix*. Using simulated population genomic data, *OptM* performs equally as well as the currently recommended threshold of 99.8% variation explained. However, when tested on empirical data of numerous populations and complex evolutionary history (where the true m is quite large), *OptM* can suggest a quantitative and producible measure of the optimal number of migration edges that best explain the tree graph. It must be noted that *OptM* is not attempting to infer the actual number of migration edges, although in less complex scenarios this is possible, but rather a reduced number that can be prioritized for their ability to best improve the fit to the tree model.

Although *OptM* is very fast, it requires multiple runs of the *Treemix* algorithm which can be computationally intensive, especially for large values of m. However, multiple iterations of *Treemix*, especially while varying the SNP-block length (-k) and/or bootstrapped across the input SNPs, can reduce the effects of spurious or weak migration edges. Furthermore, the relative height of various peaks in Δm can indicate other values for m worth exploring, although at less explanatory power than the maximum Δm [10]. The method of *OptM* does not resolve occasional shortcomings already described for ΔK (i.e., the $K = 2$ conundrum) [28] or *Treemix*, notably when migration is between closely related populations without outgroups, such as incorrect directionality, admixture in populations related to a truly admixed population, and underestimated migration weights when admixture proportions are high [2]. Nevertheless, with the ever-increasing availability of population genomic data for a variety of species, *OptM* serves a valuable purpose in providing a robust and reproducible tool for inferring the optimal number of migration events that can best explain extant levels of genetic variation from complex population histories.

Supplementary data

Supplementary data is available at *Biology Methods and Protocols* online.

Data availability

OptM v0.1.5 is currently available through CRAN (https://cran.r-project.org/web/packages/OptM) or web application (https://rfitak.shinyapps.io/OptM/; Supplementary Fig. S4). Additional computer code for generating the simulated datasets and running *OptM* is available in Supplementary File S1, and an example configuration file for the simulations can be found in Supplementary File S2. The domestic dog dataset from Vaysse *et al.* [22] is publicly available at http://dogs.genouest.org/SWEEP.dir/Supplemental.html.

Acknowledgements

I am grateful to the Duke Compute Cluster for providing the computational resources necessary for this study. I also thank A. Ochoa for commenting on earlier versions of this manuscript. There is no specific funding to report for this study.

Author contributions

R.R.F. completed all aspects of the study.

Conflict of interest

The author declares no conflict of interest.

References

1. Ellegren H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol* 2014;**29**:51–63.
2. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 2012;**8**:e1002967.
3. Teixeira MM, Barker BM. Use of population genetics to assess the ecology, evolution, and population structure of Coccidioides. *Emerg Infect Dis* 2016;**22**:1022–30.
4. von Wettberg EJB, Chang PL, Basdemir F *et al.* Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation. *Nat Commun* 2018;**9**:649.
5. Card DC, Schield DR, Adams RH *et al.* Phylogeographic and population genetic analyses reveal multiple species of Boa and independent origins of insular dwarfism. *Mol Phylogenet Evol* 2016;**102**:104–16.
6. Decker JE, McKay SD, Rolf MM *et al.* Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet* 2014;**10**:e1004254.
7. Foote AD, Vijay N, Avila-Arcos MC *et al.* Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat Commun* 2016;**7**:11693.
8. Parker HG, Dreger DL, Rimbault M *et al.* Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Rep* 2017;**19**:697–708.
9. Alberto FJ, Boyer F, Orozco-terWengel P *et al.* Convergent genomic signatures of domestication in sheep and goats. *Nat Commun* 2018;**9**:813.
10. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol* 2005;**14**:2611–20.
11. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;**155**:945–59.
12. Porras-Hurtado L, Ruiz Y, Santos C *et al.* An overview of STRUCTURE: applications, parameter settings, and supporting software. *Front Genet* 2013;**4**:98.
13. Novembre J. Pritchard, Stephens, and Donnelly on population structure. *Genetics* 2016;**204**:391–3.
14. Pritchard JK, Wen X, Falush D. 2010. Documentation for *structure* software: Version 2.3. https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/structure_doc.pdf (17 September 2021, date last accessed).
15. R Core Development Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2017. https://cran.r-project.org/ (17 September 2021, date last accessed).
16. Sonderegger DL, Wang H, Clements WH *et al.* Using SiZer to detect thresholds in ecological data. *Front Ecol Environ* 2009;**7**:190–5.
17. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds), *Second International Symposium on Information Theory*. Budapest: Akadémiai Kiadó. 1973, 267–81.
18. Palamara PF. ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. *Bioinformatics* 2016;**32**:3032–4.
19. DeGiorgio M, Jakobsson M, Rosenberg NA. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci USA* 2009;**106**:16057–62.
20. Danecek P, Auton A, Abecasis G *et al.*, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics* 2011;**27**:2156–8.
21. Lequarré AS, Andersson L, Andre C *et al.* LUPA: a European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. *Vet J* 2011;**189**:155–9.
22. Vaysse A, Ratnakumar A, Derrien T *et al.*, LUPA Consortium. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet* 2011;**7**:e1002316.
23. Purcell S, Neale B, Todd-Brown K *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.
24. Wang GD, Zhai WW, Yang HC *et al.* Out of southern East Asia: the natural history of domestic dogs across the world. *Cell Res* 2016;**26**:21–33.
25. Pilot M, Malewski T, Moura AE *et al.* On the origin of mongrels: evolutionary history of free-breeding dogs in Eurasia. *Proc Biol Sci* 2015;**282**:20152189.
26. Lee RB. *A History and Description of the Modern Dogs of Great Britain and Ireland. Sporting Division*. London: H. Cox, 1897.
27. Lee RB. *A History and Description of the Modern Dogs of Great Britain and Ireland. The Terriers*. London: H. Cox, 1903.
28. Janes JK, Miller JM, Dupuis JR *et al.* The K = 2 conundrum. *Mol Ecol* 2017;**26**:3594–602.

Supplementary File S1 for:

OptM: estimating the optimal number of migration edges on population trees using *Treemix*

Robert R. Fitak (Robert.Fitak@ucf.edu)
Department of Biology, Genomics and Bioinformatics Cluster, University of Central Florida

Methods

Mathematical derivation of Δm

The parameter $L(m)$ corresponds with the composite likelihood of the model fit using *TREEMIX* for m migration edges. $L(m)$ is thus equivalent to $L(\widehat{W}|W)$ from equation 28 in Pickrell and Pritchard (2012):

$$L(\widehat{W}|W) = \prod_{i=1}^{m}\prod_{j=1}^{m} N(\widehat{W_{ij}}|G, \hat{\sigma}^2_{ij}) \tag{1}$$

First, the mean $L(m)$ is estimated across N iterations for successive values of m where $m \geq 0$:

$$L(m) = \frac{1}{N}\sum_{i=1}^{N} L(m_i) \tag{2}$$

Then, following the methodology outlined by Evanno et al (2005), the mean difference between successive composite likelihoods, or the first order rate of change $L'(m)$, is calculated as:

$$L\prime(m) = L(m) - L(m-1) \tag{3}$$

Next, the second order rate of change in $L(m)$ with respect to m is calculated:

$$L\prime\prime(m) = |L\prime(m+1) - L\prime(m)| \tag{4}$$

or by simple algebra:

$$\begin{gathered} L\prime\prime(m) = |L(m+1) - L(m) - L(m) + L(m-1)| \\ L\prime\prime(m) = |L(m+1) - 2 * L(m) + L(m-1)| \end{gathered} \tag{5}$$

In the last step, $\varDelta m$ is calculated by normalizing $L\prime\prime(m)$ by the standard deviation in $L(m)$, $\sigma_{L(m)}$:

$$\varDelta m = \frac{L\prime\prime(m)}{\sigma_{L(m)}} \tag{6}$$

Simulated datasets

The simulated datasets were generated using ARGON v0.1 (Palamara 2016). ARGON simulates the discrete time Wright Fisher process backwards in time quickly for whole-genome sized datasets for arbitrary, user-defined demographic histories. The population graph without any migration events was identical to that in Pickrell and Pritchard (2012) which was based upon recreating patterns of diversity and linkage disequilibrium consistent with that of SNP genotype data for many modern human datasets (DeGiorgio et al. 2009) (**Fig. S1**). The parameters across simulations included:

- a serial bottleneck model composed of 20 populations
- each population had a current effective population size of 10000
- each serial bottleneck event originated from 250 individuals ever 100 generations
- a mutation rate of 10^{-8} substitutions site^{-1} generation^{-1}
- a recombination rate of 10^{-8} recombinations site^{-1} generation^{-1}
- recombinations must occur at least every 100 bp apart
- a sequence (chromosome) length of 250 megabases
- 60 chromosomes (30 diploid individuals) sampled from each population
- All populations shared a common ancestor 2000 generations in the past

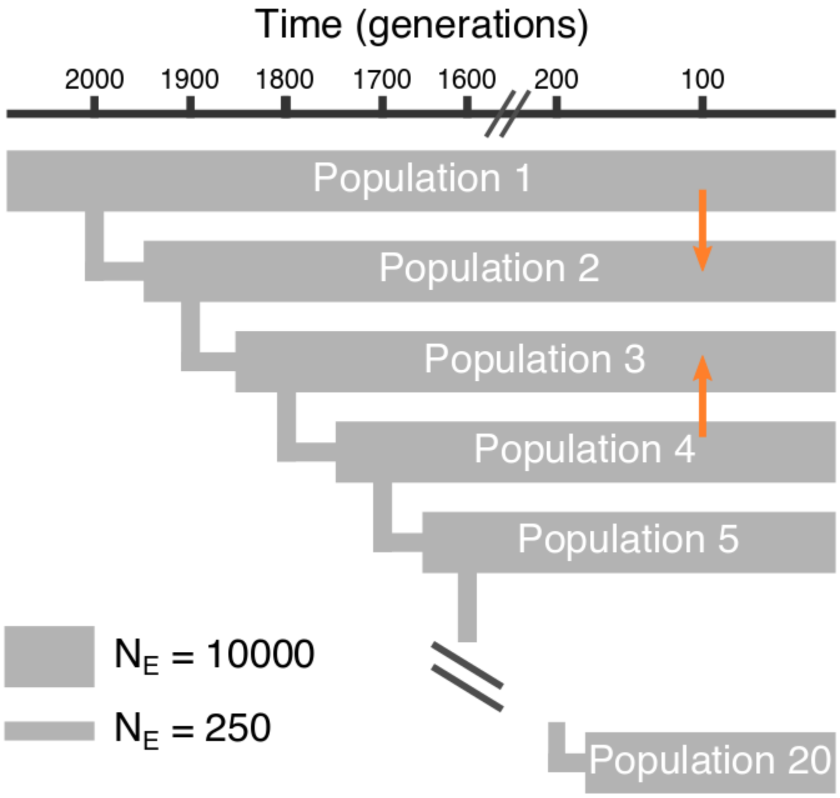


Fig. S1 Diagram of the serial population bottleneck model used for the simulated datasets. The model is as described according to Pickrell and Pritchard (2012) and (DeGiorgio et al. 2009). Two hypothetical migration edges, from population 1 into 2, and 4 into 3 are shown (orange arrows)

Either 1, 3, 5, or 8 migrations events we included in each simulation. The source and recipient population for each migration event were selected at random without replacement, and the recipient population received 30% of its genetic ancestry from the source population. All simulation events occurred 100 generations in the past. The 8 randomly selected migration edges were as follows:

1. 13 ---> 5
2. 4 ---> 12
3. 7 ---> 16
4. 6 ---> 3
5. 9 ---> 17
6. 15 ---> 1
7. 14 ---> 8
8. 19 ---> 10

An example run of ARGON v0.1 for $m = 8$. The configuration file `M8_Argon.txt` is available as a Supplementary data file.

```
# Set seed
SEED=$RANDOM
echo "$SEED"

# Run ARGON v0.1 for 8 migration edges
java -jar ARGON.0.1.jar \
   -N M8_Argon.txt \
   -pop 20 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 \
   -size 250 \
   -rec 1E-8 \
   -mut 1E-8 \
   -out M8 \
   -gz \
   -len 100 \
   -seed $SEED
```

The output from ARGON v0.1 is a compressed VCF file (e.g. above: `M8.vcf.gz`). The VCF files were converted to the PED/MAP format utilized by PLINK v1.07 (Purcell et al. 2007) with VCFTOOLS v0.1.13 (Danecek et al. 2011). This step also removed any SNP loci with a minimum allele frequency (MAF) < 0.05.

```
# Convert VCF to PED/MAP and remove maf<0.05
zcat M8.vcf.gz | \
   vcf-sort | \
   vcftools \
      --gzvcf - \
      --maf 0.05 \
      --plink \
      --out M8
```

Next, the FID and IID columns in the PLINK-formatted file were corrected so the FID column is the population (1-20), and the IID column is the individual number (1-30). A cluster file, utilized by PLINK to calculate stratified allele frequencies, was also created. The cluster file is a 3 column list of FID, IID, and the cluster (a.k.a. population) to which the sample belongs. The cluster here is the same as the FID.

```
# Change FID and IID encoding
cut -f2- M8.ped | \
   tr "_" "\t" > tmp
mv tmp M8.ped

# Make cluster file
for pop in {1..20}
   do
   for i in {1..30}
      do
      echo "$pop $i $pop" >> pops.cluster
   done
done
```

Once the PLINK PED file and the cluster file were prepared, PLINK was used to make allele counts within each cluster (population). The resulting stratified allele counts file was compressed and converted to a TREEMIX input file using the python script `plink2treemix.py`. The script plink2treemix.py is distributed along with TREEMIX v1.13 and can be downloaded from the link provided.

```
# Make a stratified allele frequency (counts) file
plink \
   --file M8 \
   --noweb \
   --freq \
   --within pops.cluster \
   --out M8

# Compress the file
gzip M8.frq.strat

# Convert to TREEMIX input file
plink2treemix.py M8.frq.strat.gz M8.treemix.gz
```

The last step was to run TREEMIX. TREEMIX was run for 10 replicates each for $m = [1..10]$. To avoid converging on the same composite likelihood for each replicate, the number of SNPs per window (`-k`) was varied across runs from 100-1000 in 50 SNP increments. A global set of rearrangements (`-global`) was also included.

```
# Run 100 runs of treemix
   # m = number of migration edges
   # i = number of replicates for each value of m
   # k = number of SNPs per window
   # s = random seed

for m in {1..10}
   do
   for i in {1..10}
      do

      # Generate random seed
      s=$RANDOM
      echo "Random seed = ${s}"

      # Generate random k between 100 and 1000 in 50 SNP increments
      k=$(seq 100 50 1000 | shuf -n 1)

      treemix \
         -i M8.treemix.gz \
         -o M8.${i}.${m} \
         -global \
         -m ${m} \
         -k ${k} \
         -seed ${s}
      done
```

```
done
```

Once completed, the TREEMIX output files were analyzed with OptM v0.1.5 using default parameters and plotted for *m* = 1, 3, 5, 8 simulated models (**Fig. S2**). All analyses were performed in R v3.4.3.

```
# Install and Load OptM v0.3
install.packages("OptM")
library(OptM)

# Load treemix output files
   # Example shown below: all treemix output for m=8 are inside a folder called "M8"
data = optM("M8")

# Plot results
plot_optM(data)
```

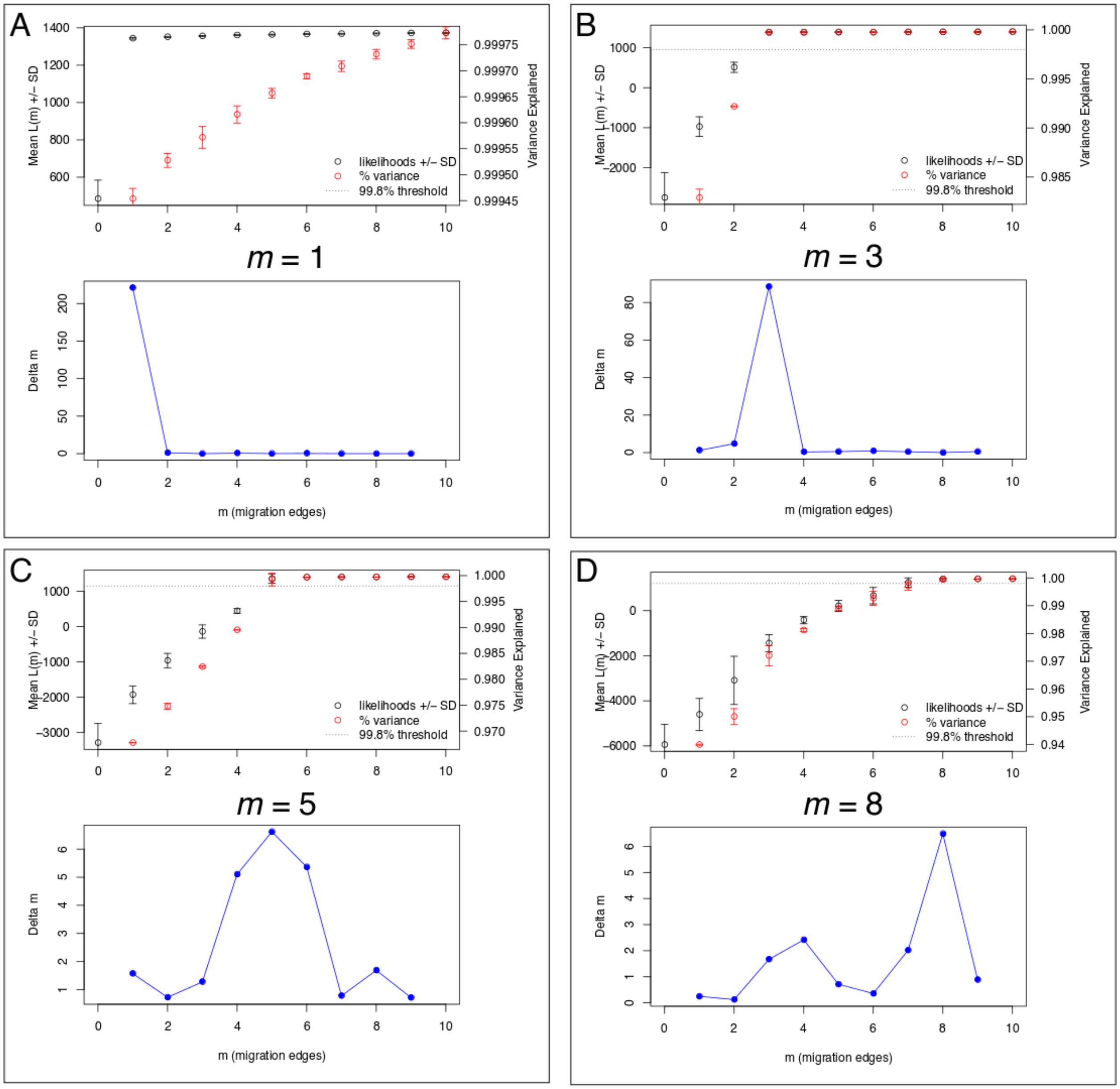


Fig. S2 The output from OptM for 1 (A), 3 (B), 5 (C), and 8 (D) simulated migration edges. The upper panels in A-D show the mean composite likelihoods (black circles, left axis) and percent variance (red circles, right axis) explained by the model with m edges. The horizontal dashed line shows the 99.8% variation threshold recommended by Pickrell and Pritchard (2012) as the cutoff for adding migration edges. In each case, OptM recovers the correct, or optimal, number of migration edges as indicated by the peak in Δm in the lower panels of A-D (blue lines). Notice that for $m = 1$, even the null model ($m = 0$) exceeds the 99.8% threshold

Empirical domestic dog and wolf dataset

As an empirical example, OptM was applied to a dataset composed of 532 domestic dogs from 48 breeds and 15 wolves genotyped on the CanineHD BeadChip (Lequarré et al. 2011; Vaysse et al. 2011). A total of 15 breeds were removed because the number of genotyped individuals was less than eight - and thus susceptible to high variance in estimates of allele frequencies. The SNPs were filtered to include only autosomal loci with a minimum allele frequency ≥ 0.05 and a genotyping rate ≥ 0.9 using PLINK. Individuals with a genotyping rate ≤ 0.9 were omitted from the analysis. Links to the dataset are shown in the code below.

```
# Download Lupa dataset, PLINK PED/MAP format
curl \
   -o lupa.map \
   http://dogs.genouest.org/SWEEP.dir/HDselection_updated_trees.map
curl \
   -o lupa.ped \
   http://dogs.genouest.org/SWEEP.dir/HDselection_updated_trees.ped

# Get sex chromosome SNPs (X = 39, Y = 40)
grep \
   -E "^39|^40" lupa.map | \
    cut -f2 > XY.exclude

# Make a list of breeds to remove: n<8
echo "ASh
Chi
CKC
CWD
DAL
ECS
ESS
FcR
Hus
LMu
Mop
Sam
Sar
Scn
Ter" > breeds-too-few.txt

# Grab the FID and IID for these breeds form the PED file



grep \
   -F \
   -f breeds-too-few.txt lupa.ped | \
   cut -d" " -f1-2 > IDs-remove.txt

# Clean data in plink
plink \
   --noweb \
   --nonfounders \
   --dog \
   --file lupa \
   --remove IDs-remove.txt \
   --exclude XY.exclude \
   --maf 0.05 \
   --geno 0.1 \
   --mind 0.1 \
   --make-bed \
   --out lupa.clean
```

Here is the output from PLINK:

```
@----------------------------------------------------------@
|        PLINK!       |     v1.07      |   10/Aug/2009     |
|----------------------------------------------------------|
|  (C) 2009 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Skipping web check... [ --noweb ]
Writing this text to log file [ lupa.clean.log ]
Analysis started: Sat Feb 23 14:06:27 2019

Options in effect:
  --noweb
  --nonfounders
  --dog
  --file lupa
  --remove IDs-remove.txt
  --exclude XY.exclude
  --maf 0.05
  --geno 0.1
  --mind 0.1
  --make-bed
  --out lupa.clean

** For gPLINK compatibility, do not use '.' in --out **
174810 (of 174810) markers to be included from [ lupa.map ]
Warning, found 547 individuals with ambiguous sex codes
Writing list of these individuals to [ lupa.clean.nosex ]
547 individuals read from [ lupa.ped ]
0 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
```

```
0 cases, 0 controls and 547 missing
0 males, 0 females, and 547 of unspecified sex
Reading list of SNPs to exclude [ XY.exclude ] ... 5744 read
Reading individuals to remove [ IDs-remove.txt ] ... 36 read
36 individuals removed with --remove option
Before frequency and genotyping pruning, there are 169066 SNPs
511 founders and 0 non-founders found
Writing list of removed individuals to [ lupa.clean.irem ]
3 of 511 individuals removed for low genotyping ( MIND > 0.1 )
Total genotyping rate in remaining individuals is 0.983518
4151 SNPs failed missingness test ( GENO > 0.1 )
28349 SNPs failed frequency test ( MAF < 0.05 )
After frequency and genotyping pruning, there are 138306 SNPs
After filtering, 0 cases, 0 controls and 508 missing
After filtering, 0 males, 0 females, and 508 of unspecified sex
Writing pedigree information to [ lupa.clean.fam ]
Writing map (extended format) information to [ lupa.clean.bim ]
Writing genotype bitfile to [ lupa.clean.bed ]
Using (default) SNP-major mode

Analysis finished: Sat Feb 23 14:07:08 2019
```

Next, a cluster file which lists each individual and the cluster (population) assignment was generated for use with PLINK. Then, the TREEMIX input file was created as described above for the simulated datasets.

```
# Clean FAM file
cut \
   -d" " \
   -f1 lupa.clean.fam | \
   sed "s/_.*//g" | \
   paste \
      -d" " \
      - \
      <(cut -d" " -f2- lupa.clean.fam) > new.fam
mv new.fam lupa.clean.fam

# Prepare cluster file
cut -d" " -f1-2 lupa.clean.fam | \
   paste -d" " - \
   <(cut -d" " -f1 lupa.clean.fam) > within.txt

# Prepare the allele counts file
plink \
   --noweb \
   --dog \
   --nonfounders \
   --bfile lupa.clean \
   --freq \
   --within within.txt \
   --out Lupa.treemix

# Compress the stratified allele counts file
gzip Lupa.treemix.frq.strat

# Download the script plink2treemix.py
wget https://bitbucket.org/nygcresearch/treemix/downloads/plink2treemix.py
chmod 770 plink2treemix.py

# Convert to a TREEMIX input file
./plink2treemix.py Lupa.treemix.frq.strat.gz Lupa.treemix.gz
```

Finally, TREEMIX v1.13 was run on the dog dataset for $m = 1 - 40$ with 10 iterations for each value of m. A window (`-k`) of 500 SNPs was used, along with a global rearrangement (`-global`). The optimal number of migration edges was estimated using OptM. Although a `for` loop is shown below, it is recommended to submit these as separate jobs to a compute cluster to reduce the time taken for analysis.

```
# Run treemix 10 times for m from 1-40 (400 runs)
   # m = number of migration edges
   # i = number of replicates for each value of m

for m in {1..40}
   do
   for i in {1..10}
      do

      # Generate random seed
      s=$RANDOM
      echo "Random seed = ${s}"

      # Run treemix
      treemix \
         -i Lupa.treemix.gz \
         -o Lupa.${i}.${m} \
         -global \
         -m ${m} \
         -k 500 \
         -seed ${s}
      done
done
```

```
# in R v3.4.3
# Load OptM
library(OptM)

# Run OptM (all files are in a folder called "LUPA")
Lupa.out = optM("LUPA")

# Plot Results
plot_optM(Lupa.out)

# Load plotting functions from Treemix
source("/Path/to/treemix-1.13/src/plotting_funcs.R")

# Plot tree with migration edges for one iteration of m=5 (Fig. 3 in main text)
plot_tree("Lupa.1.5")
```

Fitting threshold models

OptM is also integrated with the SiZer v0.1-5 package (Sonderegger et al. 2009) for fitting various ecological threshold models (see **Fig. S3 and Table S1** below for models M1, M3, M5, and M8 simulated above). These models, such as a 'piecewise linear' (PL) and 'bent cable' (BC) models, can be used to estimate the threshold, or change point, of the response as a function of an independent variable. On one side of a change point, small increases in the independent variable produce negligible effects on the response, whereas on the other side of the change point, small increases in the independent variable can produce substantial effects on the response. In the context of estimating the optimal value of m, the goal is to identify the change point where an increase in m no longer produces a worthwhile increase in $L(m)$. Both the PL and BC models are parametric models that fit two linear relationships connected by either an abrupt or a quadratic bend, respectively. OptM also fits a simple exponential model as well as a non-linear least squares exponential model primarily for visualization purposes. All parametric models are compared with the Akaike information criterion (AIC; Akaike 1973). Finally, the non-parametric “significant zero crossings” method (SiZer; Sonderegger et al. 2009), which predicts change points by examining the first and second derivatives of non-parametric smoothing functions, is also available for comparison purposes if of interest to the user but is not shown below. Please refer to the OptM manual.

```
# in R v3.4.3
# Load OptM
library(OptM)

# Load treemix output files, from simulated datasets
   # Example shown below: all treemix output for m=8 are inside a folder called "M8"
data = optM("M8", method = "linear")

# Print the model comparison (AIC) output
data$out

# Plot results
plot_optM(data, method = "linear")
```

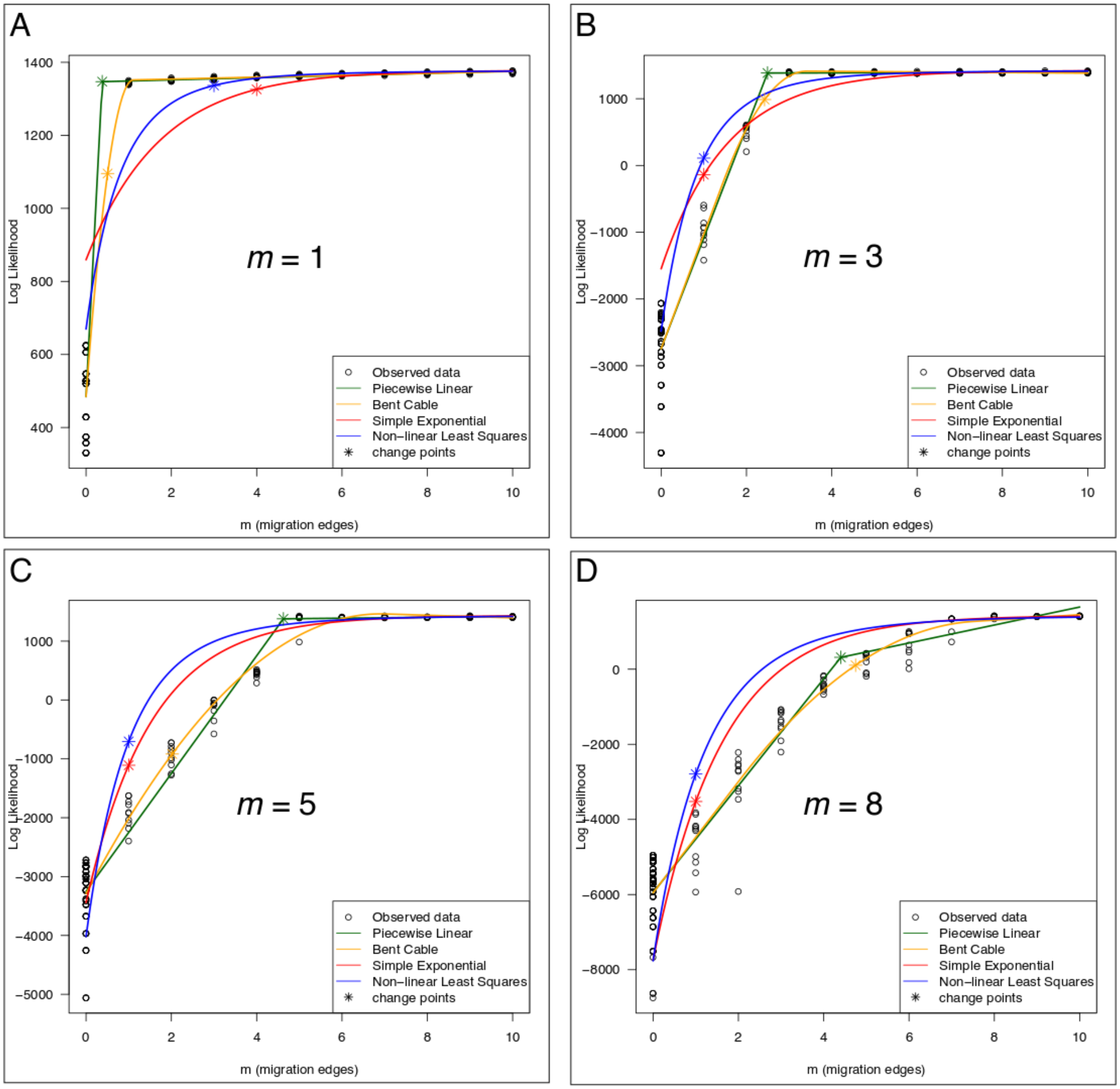


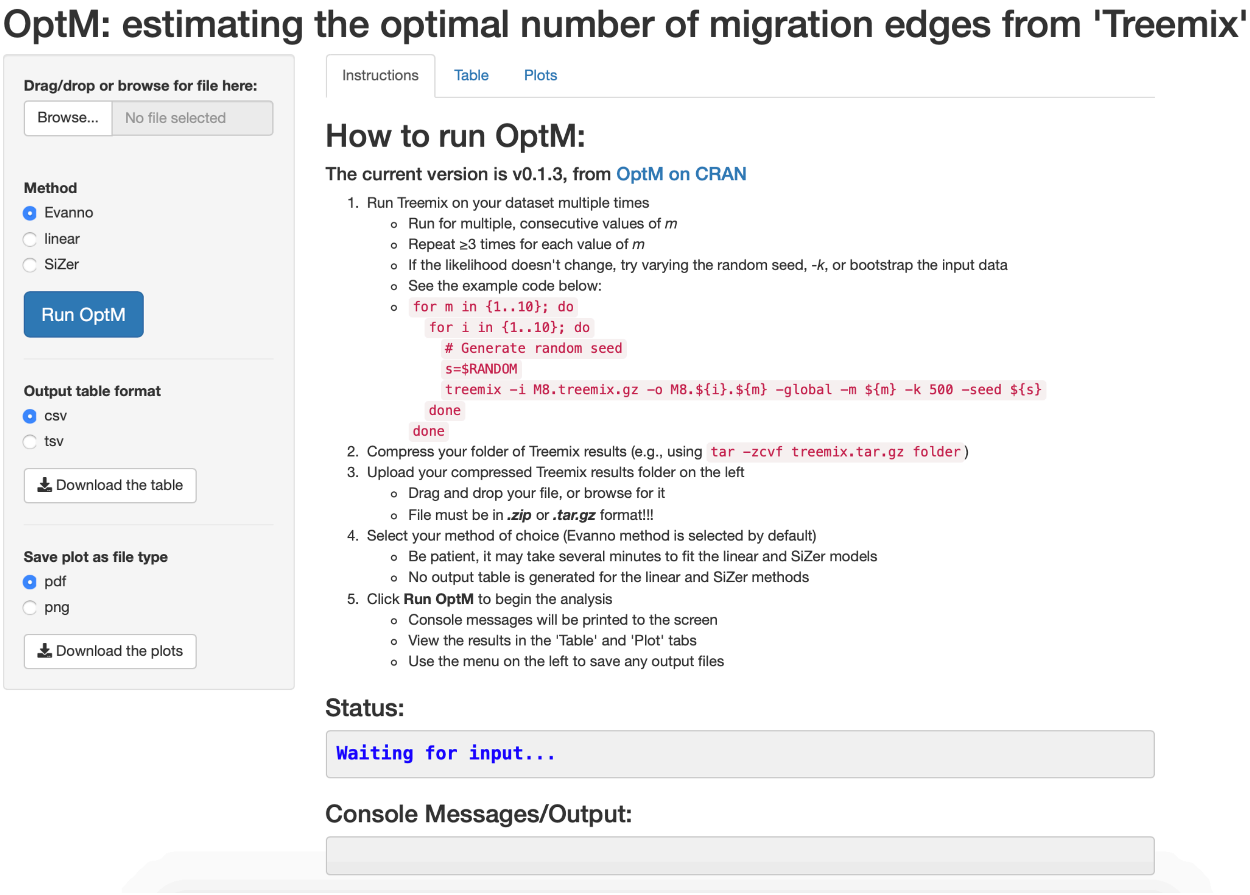
Fig. S3 The output from various ecological threshold models fit using OptM for 1 (A), 3 (B), 5 (C), and 8 (D) simulated migration edges. Each panel shows the observed composite likelihoods (black circles) for each TREEMIX run, and the four models fit (colored lines, see legend). The change points predicted by each model are shown as colored stars. Change points for the simple exponential and non-linear least squares models should not be considered accurate. For a complete description of the threshold models shown, see Sonderegger et al. 2009

Model	df	AIC	ΔAIC	Change Point
M1				
Non-linear Least Squares	3	403.6	0	3.0
Simple Exponential	3	545.9	142.3	4.0
Piecewise Linear	5	2278.0	1874.4	0.38
Bent Cable	6	2279.9	1876.3	0.51
M3				
Non-linear Least Squares	3	464.9	0	1.0
Simple Exponential	3	575.7	110.8	1.0
Piecewise Linear	5	3016.4	2551.4	2.5
Bent Cable	6	3018.0	2553.0	2.4
M5				
Simple Exponential	3	455.9	0	1.0
Non-linear Least Squares	3	456.0	0.060	1.0
Bent Cable	6	2980.5	2524.6	2.0
Piecewise Linear	5	2989.8	2533.9	4.6
M8				
Simple Exponential	3	418.0	0	1.0
Non-linear Least Squares	3	503.7	85.7	1.0
Bent Cable	6	3204.8	2786.8	4.7
Piecewise Linear	5	3206.9	2788.9	4.4

Table S1 The output from various ecological threshold models fit using OptM for 1 (M1), 3 (M3), 5 (M5), and 8 (M8) simulated migration edges. Models are ordered by the ΔAIC. df = degrees of freedom, AIC = Akaike information criterion. Change points for the simple exponential and non-linear least squares models should not be considered accurate. For a complete description of the threshold models shown, see Sonderegger et al. 2009

OptM Web Application

To make OptM as easy as possible to implement, especially for those without essential programming skills in R, we have also generated a web interface to *OptM* (**Fig. S4**; https://rfitak.shinyapps.io/OptM/). The web application allows the user to quickly load a zipped folder of TREEMIX v1.13 results, including by simple drag-and-drop, select a few options if necessary using check boxes, then run the analysis. The output from OptM will be quickly generated and viewed in separate *Table* and *Plots* tabs. The results can be downloaded in multiple formats with the simple click of a mouse.



OptM: estimating the optimal number of migration edges from 'Treemix'

Drag/drop or browse for file here:

Browse... No file selected

Method

● Evanno
○ linear
○ SiZer

Run OptM

Output table format

● csv
○ tsv

Download the table

Save plot as file type

● pdf
○ png

Download the plots

Instructions | Table | Plots

How to run OptM:

The current version is v0.1.3, from OptM on CRAN

1. Run Treemix on your dataset multiple times
   - Run for multiple, consecutive values of *m*
   - Repeat ≥3 times for each value of *m*
   - If the likelihood doesn't change, try varying the random seed, -*k*, or bootstrap the input data
   - See the example code below:
   -
```
for m in {1..10}; do
  for i in {1..10}; do
    # Generate random seed
    s=$RANDOM
    treemix -i M8.treemix.gz -o M8.${i}.${m} -global -m ${m} -k 500 -seed ${s}
  done
done
```
2. Compress your folder of Treemix results (e.g., using `tar -zcvf treemix.tar.gz folder`)
3. Upload your compressed Treemix results folder on the left
   - Drag and drop your file, or browse for it
   - File must be in ***.zip*** or ***.tar.gz*** format!!!
4. Select your method of choice (Evanno method is selected by default)
   - Be patient, it may take several minutes to fit the linear and SiZer models
   - No output table is generated for the linear and SiZer methods
5. Click **Run OptM** to begin the analysis
   - Console messages will be printed to the screen
   - View the results in the 'Table' and 'Plot' tabs
   - Use the menu on the left to save any output files

Status:

```
Waiting for input...
```

Console Messages/Output:

Fig. S4 The web implementation of *OptM* built using R shiny.

References

1. Akaike,H. (1973) Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. (Petrov BN, Csaki F, eds), pp. 267-281. Akadémiai Kiadó, Budapest.
2. Danecek,P. et al. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
3. DeGiorgio,M., et al. (2009) Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci. USA*, **106**, 16057.
4. Evanno,G. et al. (2005) Detecting the number of clusters of individuals using the software structure: a

simulation study. *Mol. Ecol.*, **14**, 2611-2620.

5. Lee,R. (1897) A history and description of the modern dogs of Great Britain and Ireland. Sporting division vol 1. H. Cox, London.
6. Lee,R. (1903) A history and description of the modern dogs of Great Britain and Ireland. The terriers vol 3. H. Cox, London.
7. Lequarré,A. et al. (2011) LUPA: A European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. *Vet. J.*, **189**, 155-159.
8. Palamara,P. (2016) ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. *Bioinformatics*, **32**, 3032-3034.
9. Parker,H., et al. (2017) Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Rep.*, **19**, 697-708.
10. Pickrell,J. and Pritchard J. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.*, **8**, e1002967.
11. Pilot,M. et al. (2015) On the origin of mongrels: evolutionary history of free-breeding dogs in Eurasia. *Proc. Biol. Sci.*, **282**, 20152189
12. Sonderegger,D. et al. (2009) Using SiZer to detect thresholds in ecological data. *Front. Ecol. Environ.*, **7**, 190-195.
13. Vaysse,A. et al. (2011) Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.*, **7**, e1002316.
14. Wang,G. et al. (2016) Out of southern East Asia: the natural history of domestic dogs across the world. *Cell Res.*, **26**, 21-33