Unicentre
CH-1015 Lausanne
http://serval.unil.ch

Year : 2014

THE USE OF SIMULATIONS IN EVOLUTIONARY POPULATION GENETICS: APPLICATIONS ON HUMANS, OWLS AND VIRTUAL ORGANISMS

KANITZ Ricardo

KANITZ Ricardo, 2014, THE USE OF SIMULATIONS IN EVOLUTIONARY POPULATION GENETICS: APPLICATIONS ON HUMANS, OWLS AND VIRTUAL ORGANISMS

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_1D7A41307FFB5

Droits d'auteur
L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright
The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Département d'écologie et évolution

THE USE OF SIMULATIONS IN EVOLUTIONARY POPULATION GENETICS: APPLICATIONS ON HUMANS, OWLS AND VIRTUAL ORGANISMS

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Ricardo KANITZ

Master en Zoologie de la Université Pontificale Catholique du Rio Grande do Sul (PUCRS)

Jury

Prof. Pierre Goloubinoff, Président
Prof. Jérôme Goudet, Directeur de thèse
Prof. Nicolas Perrin, Expert
Prof. Laurent Excoffier, Expert

Lausanne 2014

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur	Prof.	Pierre	**Goloubinoff**
Directeur de thèse	Monsieur	Prof.	Jérôme	**Goudet**
Experts	Monsieur	Prof.	Nicolas	**Perrin**
	Monsieur	Prof.	Laurent	**Excoffier**

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Ricardo Kanitz

Master en Zoologie de "Pontificia Universidade Católica do Rio Grande do Sul",
Porto Alegre, Brésil

intitulée

THE USE OF SIMULATIONS IN EVOLUTIONARY POPULATION GENETICS: APPLICATIONS ON HUMANS, OWLS AND VIRTUAL ORGANISMS

Lausanne, le 23 juillet 2014

pour La Doyenne
de la Faculté de Biologie et de Médecine

Prof. Pierre Goloubinoff

"*It is sometimes said that scientists are unromantic, that their passion to figure out robs the world of beauty and mystery. But is it not stirring to understand how the world actually works — that white light is made of colors, that color is the way we perceive the wavelengths of light, that transparent air reflects light, that in so doing it discriminates among the waves, and that the sky is blue for the same reason that the sunset is red? It does no harm to the romance of the sunset to know a little bit about it.*"

Carl Sagan

Pale Blue Dot: A Vision of the Human Future in Space (1994)

Contents

Summary 5

Résumé (en français) 6

Acknowledgments 7

Preface 9

General Introduction 10

  In silico science and evolutionary biology 10

    Model-based inference & approximate Bayesian computation 14

  Of patterns and processes 20

    Clines and clusters 22

    Selection and drift 23

  Of humans and owls (and virtual organisms) 24

Chapter 1 – A simple range-expansion model replicates the general patterns of neutral genetic diversity observed in humans 26

  Introduction 29

  Material and Methods 32

  Results 37

  Discussion 40

  Supplemental data 48

  Acknowledgements 48

  References 49

  Supporting Information 55

    Supplementary Figures 55

    Supplementary Tables 62

Chapter 2 – Natural selection in a post-glacial range expansion: the case of the colour cline in the European barn owl 66

  Abstract 68

  Introduction 69

  Material and Methods 74

  Results 87

  Discussion 91

Acknowledgements 96
Data Accessibility 97
Author Contributions 97
Supporting Information 97
References 98
Supplementary Figures 107
Supplementary Table 111
Chapter 3 – Natural selection in range expansions: insights from a spatially explicit ABC approach 114
Abstract 116
Introduction 117
Material & Methods 120
Results 126
Discussion 129
Acknowledgments 138
References 139
Supplementary Material 144
Supplementary Tables 146
Supplementary Figures 147
Supplementary References 151
General Discussion 152
Drift: the null hypothesis 154
Perspectives 155
Conclusion 159
General References 160

Summary

Computer simulations provide a practical way to address scientific questions that would be otherwise intractable. In evolutionary biology, and in population genetics in particular, the investigation of evolutionary processes frequently involves the implementation of complex models, making simulations a particularly valuable tool in the area. In this thesis work, I explored three questions involving the geographical range expansion of populations, taking advantage of spatially explicit simulations coupled with approximate Bayesian computation. First, the neutral evolutionary history of the human spread around the world was investigated, leading to a surprisingly simple model: A straightforward diffusion process of migrations from east Africa throughout a world map with homogeneous landmasses replicated to very large extent the complex patterns observed in real human populations, suggesting a more continuous (as opposed to structured) view of the distribution of modern human genetic diversity, which may play a better role as a base model for further studies. Second, the postglacial evolution of the European barn owl, with the formation of a remarkable coat-color cline, was inspected with two rounds of simulations: (i) determine the demographic background history and (ii) test the probability of a phenotypic cline, like the one observed in the natural populations, to appear without natural selection. We verified that the modern barn owl population originated from a single Iberian refugium and that they formed their color cline, not due to neutral evolution, but with the necessary participation of selection. The third and last part of this thesis refers to a simulation-only study inspired by the barn owl case above. In this chapter, we showed that selection is, indeed, effective during range expansions and that it leaves a distinguished signature, which can then be used to detect and measure natural selection in range-expanding populations.

Résumé (en français)

Les simulations fournissent un moyen pratique pour répondre à des questions scientifiques qui seraient inabordable autrement. En génétique des populations, l'étude des processus évolutifs implique souvent la mise en œuvre de modèles complexes, et les simulations sont un outil particulièrement précieux dans ce domaine. Dans cette thèse, j'ai exploré trois questions en utilisant des simulations spatialement explicites dans un cadre de calculs Bayésiens approximés (approximate Bayesian computation : ABC). Tout d'abord, l'histoire de la colonisation humaine mondiale et de l'évolution de parties neutres du génome a été étudiée grâce à un modèle étonnement simple. Un processus de diffusion des migrants de l'Afrique orientale à travers un monde avec des masses terrestres homogènes a reproduit, dans une très large mesure, les signatures génétiques complexes observées dans les populations humaines réelles. Un tel modèle continu (opposé à un modèle structuré en populations) pourrait être très utile comme modèle de base dans l'étude de génétique humaine à l'avenir. Deuxièmement, l'évolution postglaciaire d'un gradient de couleur chez l'Effraie des clocher (*Tyto alba*) Européenne, a été examiné avec deux séries de simulations pour : (i) déterminer l'histoire démographique de base et (ii) tester la probabilité qu'un gradient phénotypique, tel qu'observé dans les populations naturelles puisse apparaître sans sélection naturelle. Nous avons montré que la population actuelle des chouettes est sortie d'un unique refuge ibérique et que le gradient de couleur ne peux pas s'être formé de manière neutre (sans l'action de la sélection naturelle). La troisième partie de cette thèse se réfère à une étude par simulations inspirée par l'étude de l'Effraie. Dans ce dernier chapitre, nous avons montré que la sélection est, en effet, aussi efficace dans les cas d'expansion d'aire de distribution et qu'elle laisse une signature unique, qui peut être utilisée pour la détecter et estimer sa force.

Acknowledgments

I would like to express my gratitude towards many people and institutions that allowed me to complete my studies to the level of a PhD. This achievement is the result of the encouragement provided by my parents **Sonia Ingrid Kanitz** and **Walmor Ari Kanitz** and the opportunities resulting from this encouragement. An incentive also provided to my dearest sister, and soon PhD as well, **Ana Carolina Kanitz**. Thanks to my family, I could study in a very good school (Colégio Sinodal) and enter at the best Biology BSc in Brazil, my *alma mater*, Universidade Federal do Rio Grande do Sul (**UFRGS**). I am very thankful to the many good teachers I had in these two institutions, who thought me a great deal about, not only life sciences, but about life itself.

In the beginning of my scientific career, I had the wonderful opportunity to work alongside two fantastic researchers at Pontifícia Universidade Católica do Rio Grande do Sul (**PUCRS**): Prof. **Sandro L. Bonatto** and Prof. **Nelson J. R. Fagundes**. Thank you and see you soon in Brazil!

In 2010, when I was looking for good PhD-student position, I was incredibly lucky to come across something much better than just good. I would like to thank Prof. **Jérôme Goudet** for accepting me in his team, for trusting in my work and for his guidance in these last four years. It has been a pleasure to spend this time at the University of Lausanne (**UNIL**), and its campus with what I am sure is the best view in the world. In our group, I could count with precious collaboration, help and advice of Dr. **Samuel Neuenschwander** and **Sylvain Antoniazza**. Especial thanks to you guys for being there!

During my stay in Switzerland I was fortunate to meet many good people and make several good friends. I would like to thank these people for playing this vital role in my life of a social ape, making my life always a bit psychologically healthier. So, thanks to my former and current office mates: **Olivier**, **Katie**, **Mikko**, **Alok**, **Elisa**, **Débora**, **Emanuelle** and **Samuel** (again). Also, my faithful lunch, coffee and beer-break mates during these four years: **Fardo**, **Valentijn**, **Dumas**, **Ivan**, **Erica**, **Miguel**, **Slimane**, **Manuel**, **Simon**, **Matthias**, **Anshu** and **Lucas**. Also, many other people from the DEE for sharing nice conversations, helping out with specific question or just for very enthusiastic “good mornings”. This goes especially for **Nicolas**, **Anna**, **Eric**, **Anahí**, **Nadja**, **Martha**, **Tomasz**, **Pawel**, **Nils**, **Marie**, **Marie**, **Guillaume**, **Arnaud**, **Christophe**, **Fabrice**, **Guangpeng,** Prof. **Pannell**, Prof. **Perrin**, and Dr. **Fumagalli**.

Finally, my biggest thanks go to my love, my partner and witness in life, **Aline Xavier da Silveira dos Santos**. Without you, I would not have managed. This thesis is dedicated to you!

Preface

This thesis work came into being during the past four years as a result of an "evolutionary process" (which does not necessarily imply progress) of creation (ironically). The original idea was to explore the evolution of skin color in human populations. It turned out that we did explore evolution of color, but not in humans; and we did study human evolution, but not on skin color. In fact, this thesis work has spread much further than anticipated. Instead of focusing on one question, we extended it to a myriad of problems in evolutionary biology. We looked into models of the neutral evolution of modern human populations and how this would have implications on how we deal with races in our species. We also looked into one of the oldest dilemmas in evolution: neutrality *vs.* natural selection, demonstrated that selection has happened in barn owls and that it can be assessed in essentially any other system of range expansion.

Across the whole text of this thesis report I digress about questions in evolutionary biology, but generally these questions fall within the narrower scope of population genetics and, occasionally, even phylogeography. These terms are at times applied interchangeably, but I hope the contexts in which they are presented are clear enough to identify at which levels the contributions are made. So, even though this is a work of population genetics, I trust it has implications for evolutionary biology and potential applications to phylogeography. Also, at times, I make use of the singular form of the first person (i.e. "I") to express my own particular view on a given subject. Some other times, I make use of the plural form ("we") for statements that are derived from a group work or idea. So occasional changes in the singular or plural forms are not mistakes, but are by design and do have a meaning in the context of this thesis report.

General Introduction

In silico science and evolutionary biology

Simulations can be defined as procedures used to imitate real-world systems or processes (Banks et al. 2005). Even though simulations can be used to look at virtually any sort of scientific question, they are particularly useful for studying phenomena that would be otherwise intangible due to cost, complexity, space or time constraints. Furthermore, simulations can normally be run with a large number of replicates, taking advantage of the three-century-old idea behind the law of large numbers (Bernoulli 1713; Haigh 2012). This law states that, in a survey to assess the mean value of a given trait in a population, one observes more fluctuations when the number of observations is small; but, as one increases the number of measurements, the calculated mean invariably converges to the true mean of the population. In summary, a large number of measurements lead to increased accuracy. BOX 1 provides and example of a simulation with a simple underlying model applied to the estimation of the irrational number π that can be done manually. In BOX 2, the same model can be run in R, increasing the potential number of replicates, leading to a better estimate of π.

Every simulation requires an underlying model – i.e. a logical description of how the system of interest works. In fact, a simulation is nothing more than the implementation of such model, and the quality of the simulations will eventually depend on how good the model in use is. A good scientific model, in general, is one that is able to describe as many parameters as possible with as little complexity as needed, a concept broadly known as Occam’s razor (Domingos 1999). This idea of simplicity permeates almost every simulation-based study. All models are incomplete. Therefore, no model is fully correct, and by logical extension all models are essentially wrong.

However, some – hopefully most – can be effectively used to understand the system under examination. In the famous words of George E. P. Box, "[...] *essentially, all models are wrong, but some are useful*" (Box and Draper 1987).

The first computer simulations appeared with the arrival of the very first fully programmable electronic computer (the Electronic Numerical Integrator And Computer, or simply ENIAC) in the late 1940's (Winsberg 2010). The ENIAC was first conceived by the United States Army to calculate artillery tables. These tables used to be calculated by women, who curiously were then known as the "computers". The army needed a faster and more reliable source of these calculations, leading to the expansive development of the electronic computing machine. ENIAC's first application, however, was not to calculate ballistic trajectories: When the mathematician John Von Neumann (Los Alamos National Laboratory) learned of its development, Los Alamos joined the army's engineering endeavor and redirected the efforts towards simulating a model of a thermonuclear reaction (Metropolis 1987). These simulations proved successful both on ENIAC's computation capability and the theoretical possibility of the hydrogen bomb. The calculations were performed using the Monte Carlo method: an approach that involves the repeated random sampling of values for the parameter in question exploring the parameter space of a predetermined model (MacKay 1998), evoking the law of large numbers once again (Haigh 2012). Since then, the use of simulations has burgeoned and extended to many fields of science, with especial importance in meteorology, astrophysics, economics, fluid mechanics, engineering, ecology (Winsberg 2009) and, of course, evolutionary biology (Hoban et al. 2011).

BOX 1. Running ‘simulations’ without a computer and estimating the value of π

One can execute simulation-like experiments without a computing machine. In the pre-computer era, these experiments used to be done manually. Perhaps the most famous among them was **Buffon’s needle problem** (Aigner and Ziegler 2001). It consisted in investigating, on a striped surface, what is the probability of a needle to cross the boundaries between stripes (Fig. I). As trivial as it might seem, this experiment can actually be used to approximate the value of the irrational number π. The underlying model of these ‘simulations‘ states that the probability P of a given needle, with length (L) shorter than stripes’ width (W) to cross the stripes’ boundaries is $P = N_n/n \sim 2L/\pi$. Therefore π can be approximated with many needle tosses to $\pi \sim 2Ln/N_n$, where N_n is the number of observed crossings, n is the total number of throws and L is the length of the needle relative to the stripe width. Note that L here is scaled to width, so that length is given as a fraction of the stripes’ width.

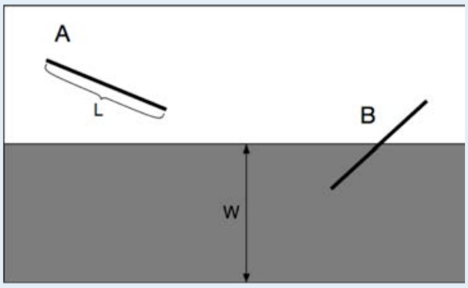

Figure 1: Buffon's needle problem illustration, where needle A does not cross the boundaries and B does. This probability depends on the relative size between stripe width (W) and needle’s length (L) and it can be used to approximate the value of π.

In evolutionary genetics, Alex S. Fraser and James Stuart F. Barker presented the earliest verifiable simulation studies in a series of eight articles from 1957 to 1960 (Fraser 1957b, a; Barker 1958b, a; Fraser 1958, 1959b, a, 1960). In Fraser (1958), the author introduces a Monte Carlo approach to simulate the effect of selection on six loci with a predetermined recombination scheme. He then compared the changes in allele frequencies under small and large population sizes and high and low intensities of selection. He observed, as expected from previous theoretical and experimental work, that higher linkage, low selection and small population size decreased the pace of adaptation of the analyzed population, making the point that simulations could, already then, be used to study evolutionary questions. Further analyses of the effect of linkage

(Fraser 1957b), autosomal (Barker 1958a) and sex-linked loci (Barker 1958b), epistasis (Fraser 1959b, a), and population structure (Fraser 1960) were presented in the subsequent papers in the series. Numerous studies followed the seminal work of Fraser and Barker in a rather continuous pace – e.g. (Gill 1964; Felsenstein 1976; Davis and Brinks 1983; Weir and Cockerham 1984) – progressively exploring different aspects of biological phenomena such as random mating, natural selection, genetic drift, genetic linkage, etc. However, no major increase in the popularity of the use of simulations was observed until the 1990's, probably due to the lack of computational power to study more complex questions, after the most straightforward ones had already been explored.

BOX 2. Buffon's needle in R

An example of the 'simulation' approach presented in Box 1 can quickly be run in R (R Core Team 2012) with the package 'animation' by executing the code below. This can also be used to demonstrate the effect of the law of large numbers. As the number of observation increases, the estimated π value gets closer to its actual value (3.14159265359, here with ten decimals).

```
library(animation)

ani.options(nmax = 200, interval = 0.1)

par(mar = c(3, 2.5, 0.5, 0.2), pch = 20, mgp = c(1.5, 0.5, 0))

buffon.needle(mat = matrix(c(1, 2, 1, 3), 2))
```

Much more recently with the popularization of personal computers and the increase of their calculation capability, a boom in the use of simulations in population genetics came about with several concomitant works including Hudson (1991) who investigated the implementation of intermediate levels of recombination – that could not be solved analytically – in a coalescent model (Kingman 1982). His approach was largely based on the Gillespie's algorithm (Gillespie 1976), which simulates continuous-time Poisson processes for parameter value assessment (Wakeley 2008). Nearly simultaneously, Bowcock et al. (1991) used simulations to draw an F_{ST} null distribution to be compared with their observed data; an approach widely used in many current studies, including the one presented in the second chapter of this thesis. After then, many other works have taken advantage of simulations to investigate various questions in the field – e.g. (Charlesworth et al. 1993; Burger and Lande 1994; Charlesworth et al. 1995; Hardy and Vekemans 1999; Balloux et al. 2000; Edmonds et al. 2004; Evanno et al. 2005; Klopfstein et al. 2006; Fagundes et al. 2007; Excoffier and Ray 2008; Peischl et al. 2013). Furthermore, a series of programs and packages for population genetics simulations have been developed in the last decade or so, as carefully reviewed in Hoban et al. (2011), with especial attention to the most complete simulator according to these authors, and the one used throughout this thesis: quantiNEMO (Neuenschwander et al. 2008a).

Model-based inference & approximate Bayesian computation

As mentioned above, the first simulations run in an electronic computer were used to explore a parameter space with the Monte Carlo algorithm (Metropolis 1987). This method, envisioned by Stanislaw Ulam and Nicholas Metropolis (Cahn 2001), consists in repeatedly sampling random values to obtain probability distributions for certain parameters. In the case of the atomic fusion reaction, put in a very simplified version, the parameter in question was the frequency of collisions between moving atomic nuclei, where the random numbers were applied to deciding which was the next move of each one of the particles in the system (Cahn 2001). When the nuclei touched, a fusion reaction would take place. In this search method, each step is independent of previous movements, following a completely random path with variable distances across the parameter space. So, essentially, there is no way in which one could guide the search towards a more likely parameter combination. A very popular example of the implementation of this search method also consists of calculating the value of the irrational number π, based on sampling random coordinates in a two-dimensional system consisting of a circle enclosed by a square, as presented in BOX 3.

BOX 3. A simple Monte Carlo approach to estimate the value of π in R

Here, we have a different and more straightforward way to estimated the value of π using a Monte Carlo approach in R. With 100’000 replicates (N), we estimate the area of a circle based on each replicate’s coordinates falling inside or outside a circle drawn inside a square with side 2. Because the area of circle is defined as $\pi \times r^2$, and here r=1, the ratio of the area of the circle over the area of the square should equal π/4.

```
N <- 100000;

x <- runif(N, min= -1, max= 1); y <- runif(N, min= -1, max= 1)

is.inside <- (x^2 + y^2) <= 1^2

(pi.estimate <- 4 * sum(is.inside) / N)

 [1] 3.1478    # Not so bad an estimate!

plot(x[ is.inside], y[ is.inside], pch = '.', col = "blue")

points(x[!is.inside], y[!is.inside], pch = '.', col = "red")
```

Since the Monte Carlo method’s elaboration, other model-based statistical inference methods have been developed, but in general they apply the same rationale of sampling a parameter space to evaluate probabilities. It is not the goal of this text to explore and explain them all, but some information is provided on a few that are particularly important in the context of evolutionary biology studies. The so-called Markov Chain Monte Carlo (MCMC) method has been widely used in phylogenetics and it consists on sampling from probability distributions with simulations, where each new step depends on the present one, but not on the past: this is a Markov chain (Hastings 1970). The idea here is essentially that in the century-old random-walk problem (Pearson 1905), where an exploratory path is taken across the parameter space by simply picking a random direction at every discrete step (Spitzer 1964). When applied to phylogenetics, MCMC’s new steps are actually slightly different phylogenetic trees, but these steps can be used in any other system where a given parameter in a simulation is modified, exploring a new position in the parameter space of interest. Another key feature of MCMC methods is that they are self-improving. Not all new steps are necessarily accepted; only the ones that increase the fit of the data to the model (the model's likelihood) are. So that MCMC runs tend to maximize the likelihood of the parameter combinations at each new step until a steady state is reached, where the variation in the parameter values do not affect anymore the overall probability of the tested model. This idea is in the very heart of Maximum Likelihood Estimation (MLE) methods (Scholz 2004). Put very simplistically, Bayesian MCMC methods, such as the ones applied in the widely used programs MrBayes (Ronquist et al. 2012) and BEAST (Drummond et al. 2012), use the exact same procedure, but limit the parameter space to be explored to so-called prior distributions (Huelsenbeck et al. 2001). So in the Bayesian approach, the final probability of the model also depends on these previously defined prior distributions. These resulting distributions are then called posterior-probability distributions, instead of maximum-likelihood distributions in MLE.

Whichever flavor of MCMC used, however, a likelihood value must be calculated, either exactly, or approximately. This is not always feasible for complex models (Beaumont et al. 2002), especially when it comes to testing evolutionary scenarios. When the likelihood of a model cannot be assessed, evolutionary biologists have applied approximate Bayesian computation (ABC, see example in BOX 4) approaches for parameter estimations (Sunnaker et al. 2013). Even though MCMC-like approaches have been developed (Wegmann et al. 2009), the way that has become traditional for implementing ABC consists of the following steps (Fig. 1):

1. **Sampling –** A very large number of simulations (e.g. 1 million) are run for as many models as one is interested in testing. Every simulation uses parameter values taken from the prior distributions, so that each simulation has a potentially unique combination of these parameters, covering the entire so-called parameter space. The more simulations are run, the more refined is the exploration of this space. And, of course, every parameter forms a new dimension in this space (e.g. a study investigating population size, migration, mutation and growth rates has four dimensions to explore). The simulations generate, as output, summary statistics whose purpose is to condense the information generated. Choosing these summary statistics is a very important step in an ABC approach. They should be able to capture all the information deriving from the varying choice of parameters, but should also be limited to the smallest possible number, since every new statistics also generates noise along with the extra information it provides. This sampling phase, therefore, consists in sampling simulations and retaining the parameter values and the summary statistics produced by them. ABC can be considered as a brute-force approach towards the exploration of the parameter space.

2. **Estimation –** The estimates generated by an ABC analysis are essentially the result of the comparison between the observed summary statistics (calculated from a real population) and the simulated summary statistics. Since statistics are connected to parameter values in the simulations, one can use this comparison to obtain a posterior probability distribution for the parameters of interest. This can be done by simply defining an interval around the observed statistics for retaining simulations out of which to extract the posteriors (rejection method), or by improving the rejection approach by implementing a

local linear regression (for the interval previously defined) to project the simulated parameter value to the position in the statistics axis where the observed statistics are (local linear regression method). This latter can also be further incremented by using weighted contribution of the different simulations according to their distance (in the statistics axis) to the observed statistics values: the closer they are, the higher their weight (weighted local linear regression method) (Beaumont et al. 2002).

3. **Validation** – This whole estimation procedure described above can be tested using the parameter values in the simulations themselves to try and re-estimate their values through the estimation step. Since the parameter values in the simulations are known, one can assess the precision and accuracy of the estimates by comparing these so-called pseudo-observations with the estimates. The most traditional and straightforward way to do so is by means of the coefficient of determination, or R^2 (Neuenschwander et al. 2008b).

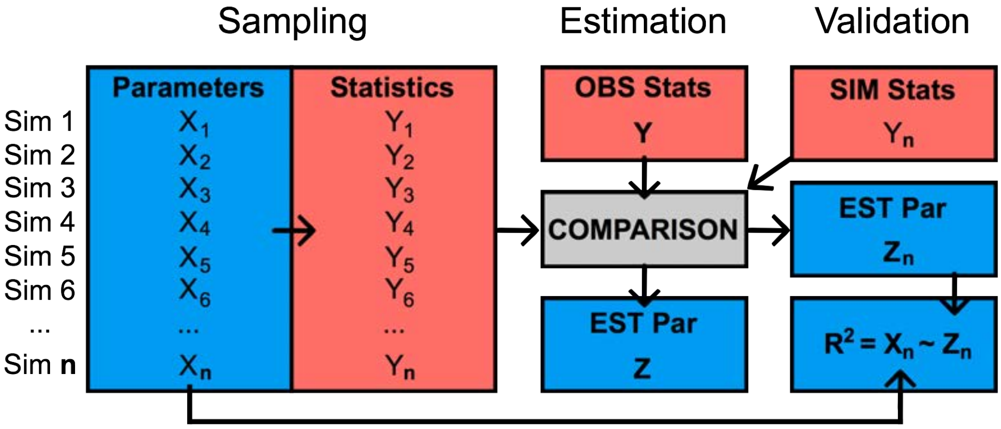

Figure 1: Schematic representation of the ABC approach for parameter estimation. The sampling, estimation, and validation steps are depicted, where **n** simulations are run with **n** parameter values ($\mathbf{X_n}$), producing **n** summary statistics values ($\mathbf{Y_n}$) that are then compared with the observed statistics (**Y**) to generate the parameter estimate **Z**. The simulations are then used to assess the quality of the estimates: the statistics trey produce ($\mathbf{Y_n}$) are used as pseudo-observed summary statistics, leading to estimates ($\mathbf{Z_n}$) that can be then compared to the pseudo-observed parameter values ($\mathbf{X_n}$). The better the fit of $\mathbf{X_n}$ and $\mathbf{Z_n}$ ($\mathbf{R^2}$), the better the quality of the estimates of that given parameter.

BOX 4. A simple demonstration of Approximate Bayesian computation (ABC) using R

As a toy example, consider a horizontal rectangle defined by **sides X** and **Y** (Fig. 1A), but out of which one can only measure **diagonal** (**D**) and **area** (**A**). The question here is: What are the values of X and Y given D and A? Even though this question can easily be solved analytically, let us try and deal with it an ABC framework. In this case **X** and **Y** are the **model parameters**; **D** and **A** are the **summary statistics**. So, the observed summary statistics of our rectangle are **A = 60** and **D = 13**. Also, being horizontal, this rectangle has a larger base length than height (i.e. X > Y) and we can also assume that X is never larger than 20 (X =< 20) as a prior of our model. To estimate the values of X and Y (Fig. 1B), here follows a simple implementation in R:

```
# Load the necessary library (install.library("abc") to install it):
library(abc)

# The OBSERVED values for A and D are:
OBS <- c(60,13)

# Generate simulated data table (10'000 simulations):
SIM <- data.frame(matrix(ncol=4,nrow=10000))
names(SIM) <- c("X","Y","A","D")
for(i in 1:dim(SIM)[1]){
  X <- runif(n=1,min=0,max=20) # Prior distr. for X
  Y <- runif(n=1,min=0,max=X) # Prior distr. for Y
  A <- X*Y
  D <- sqrt(X^2+Y^2)
  SIM[i,] <- c(X,Y,A,D)
}

# Standardize the summary statistics (for OBS and SIM):
  OBS[1] <- OBS[1]/max(SIM[,3]); OBS[2] <- OBS[2]/max(SIM[,4])
  SIM[,3] <- SIM[,3]/max(SIM[,3]); SIM[,4] <- SIM[,4]/max(SIM[,4])

# Estimate X and Y (local linear regression method):
epsilon <- 0.05 # Proportion of retained simulations
EST_X <-abc(OBS,SIM$X,SIM[,3:4],tol=epsilon,method="loclinear")$adj.values
EST_Y <-abc(OBS,SIM$Y,SIM[,3:4],tol=epsilon,method="loclinear")$adj.values

# Point estimates for X and Y (mode of posterior distr.):
EST_X[which.max(density(EST_X)$y)]
[1] 11.44782 # Not bad! The analytical solution is X = 12
EST_Y[which.max(density(EST_Y)$y)]
[1] 5.222932 # And Y = 5
```

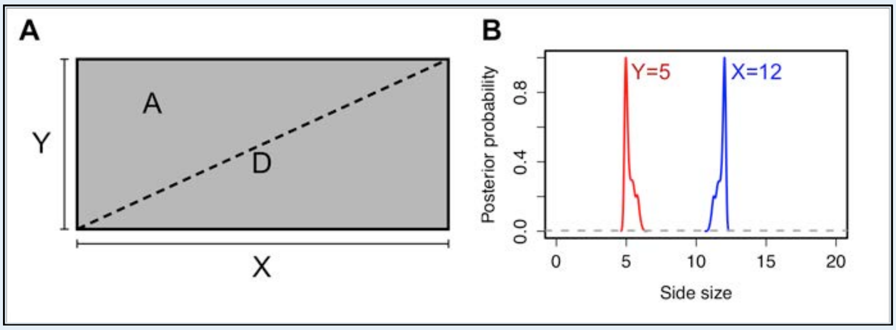

Figure 1: In **A**, the parameters and summary statistics: Sides X and Y are the parameters to be estimated; Area (A, gray rectangle) and diagonal (D, dashed line) are the summary statistics. In **B**, the posterior distributions of both **X** (blue) and **Y** (red), with the prior distribution for both parameters as the gray dashed line.

Simulation-based approaches are not only used to estimate parameter values, but also to compare competing models. In evolutionary biology, these models normally consist of different evolutionary scenarios to be tested, as competing hypotheses to explain the system under examination. There are various ways in which this comparison can be done – e.g. likelihood ratio tests, Bayes factors, etc. – but all of them are based on verifying the match between model and data. The model that shows the better match is the one chosen. This match is assessed in different ways depending on the model, but in the most used method in this thesis (ABC), this is done via the Euclidian distances measured between the statistics observed in the real dataset and the statistics produced by a given subset of best simulations of each model. The model that produces statistics that are closer to the observation is normally the chosen one. Complications are foreseeable, though. This choice will depend on the statistics chosen, the number of simulations retained and how these simulations' contribution is weighed, as well. There is an already vast literature on model-comparison approach raising some more possible issues (Templeton 2009) and solutions (Beaumont et al. 2010; Bertorelle et al. 2010). Furthermore, a validation procedure can also be run for the model comparison to evaluate possible biases and the precision of the model assignment using simulated data, as done in chapter 2 of this thesis.

Of patterns and processes

The distinction between *patterns* and *processes* is a very important notion for the understanding of biological evolution. Although introduced and popularized in a macroevolutionary context (Eldredge and Cracraft 1980), this idea has implications in all areas of evolutionary biology (Chapleau et al. 1988). In the original definition, by Eldredge and Cracraft (1980), patterns are "*aspects of the apparent orderliness of life*"; and processes are "*the mechanisms that generate these patterns*". Essentially, what they mean is that patterns are the snapshot signatures found in natural populations (horizontal in time), and processes are the paths that lead to these signatures (vertical in time). As examples of patterns, one can cite genetic structures, differential amounts of genetic diversity, allele frequency clines and other types of gradients, specific genomic signatures like runs of homozygosity, among many others. The processes in population genetics are, in their essence, combinations and modifications of the evolutionary forces (i.e. drift, selection, migration and mutation): demographic and range expansions, adaptation events, bottlenecks, reproductive isolation, secondary contacts, isolation by distance, etc. All these processes leave signatures in the form of the above-mentioned patterns.

Now, let us consider a simple example of the *process* of a bottleneck, which generates a *pattern* of low genetic diversity in the population under analysis. In this toy example, by looking at 20 microsatellite loci with a very low number of alleles (**k**) each, an imaginary evolutionary biologist – who is ignorant of the process in question – could infer that it is actually a bottleneck. She could be right, but other processes may also lead to low genetic diversity patterns (e.g. a selective sweep, a historical small population size, etc.), making our imaginary colleague’s interpretation potentially mistaken. In fact, many other patterns found in natural populations may arise due to different processes. Often, they can be told apart by including other signature patterns in the examination. In this case, if the biologist had included in her analysis the exploration of the range of allele sizes (**r**) present in these loci, she could have grasped a bit more of information about the processes being a bottleneck or not. This is because one expect **r** to decrease more slowly than **k** after such a demographic event – this is the so-called Garza-Williamson’s M statistics ($M = k/r$, (Garza and Williamson 2001)). Many studies in population genetic-related fields have been carried out with this sort of approach of including different statistics that describe patterns to infer the processes behind them. More recently, however, a movement towards a more hypothesis-driven methodology has established in the field (Knowles 2009), taking advantage of the simulation concepts presented above – with especial emphasis on ABC methods (Beaumont et al. 2002), as presented in the previous section. As to what this thesis is concerned about, all chapters here are studies focusing in deciphering the processes behind observed patterns. This is done with simulations that try to mimic these processes and that generate sets of patterns (in the form of various kinds of summary statistics) that are compared again with the observed patterns.

Clines and clusters

One debate in which this distinction between patterns and processes does not seem to be completely clear is on the distribution of genetic diversity in modern humans: the hereafter called clines-*vs.-clusters* dilemma. Most studies have focused on the patterns alone, but little attention has been given to the processes that may have lead to them (Rosenberg et al. 2002; Serre and Pääbo 2004). Here, in chapter 1, we propose a model of a simple process of demographic diffusion across a uniform environment: a model that reflects, in its essence, a continuous process. Interestingly, such model produces patterns that are very similar to the patterns observed in real human populations. On the one hand, one would expect this model to reproduce the clinal signatures already observed in the literature (Handley et al. 2007). And it does. On the other hand, surprisingly, our model also recovers, to a large extent, the cluster patterns described by others (Rosenberg et al. 2002). As a result, we believe that, even though there is a discussion on what are the most relevant patterns in human diversity, the underlying process is essentially continuous. We also make the case that this process should be considered in further studies looking for associations and signals of selection in the human genome because the proper appreciation of the background demography is vital to define null hypothesis for these studies.

Selection and drift

Occasionally, the processes of natural selection and genetic drift may lead to the same observed pattern in a population. This happens because some demographic phenomena alter genetic variation in ways that resemble selection. For instance, the neutrality test devised by Tajima (1989) checks for either the excess or the lack of low-frequency variants, which then respectively suggest the action of either purifying or balancing selection. Nevertheless, demographic expansions or reductions can also leave the same sort of signatures. That is because the changes in effective population size would either allow for the accommodation of new variants in the population or lead to an accelerated loss of variation because drift would have become stronger. In fact, the statistic derived from Tajima's test (Tajima's D) has been widely used as a demographic-variation indicator for molecular markers assumed to evolve neutrally (e.g. (Fagundes et al. 2008)).

When it comes to phenotypic variation, even more confounding factors may play a role. Phenotypes vary according to their underlying genetic, but also their environmental circumstances. To rule out acclimation to the different environments, experimental procedures – like common garden experiments (Molles and Cahill 1999) – are normally advised. Besides, a phenotypic difference does not necessarily imply in acclimation or adaptation. It can be simply neutral, as well. In gradients of variation (i.e. clines), selection used to be considered as the cause of the differentiation (e.g. (Hewitt 1996)). More recently, however, it has been demonstrated that these clines can also emerge out of purely neutral scenarios (Edmonds et al. 2004; Klopfstein et al. 2006). This would happen during range expansions, where the founder events of subsequent colonization of new areas would amplify random drift and generate potentially very differentiated genetic (and phenotypic) compositions at the starting and ending points of the expansion. So, even very coherent patterns – such as clines – may come to existence via processes as different as natural selection and random genetic drift.

Of humans and owls (and virtual organisms)

In this thesis report, I present three studies – distributed in three chapters – that cover different aspects of the problems presented above. In the first part, a simple model for the human colonization of the globe is presented, showing that a purely continuous process can lead to both clinal and cluster signatures observed in modern human populations. The second chapter, a collaborative work with Sylvain Antoniazza, approaches the post-glacial evolutionary history of the European barn owl and the emergence of a striking color cline across the continent. There we show that this cline cannot be the result of purely neutral processes, and that natural selection has to be invoked to explain its appearance and maintenance. In the last part chapter, I devise a method to measure selection strength in range-expansion scenarios, widely inspire by the barn owl case. There, I demonstrate that selection is actually effective in this drift-prone situation, and that it leaves consistent signature that can be used to assess its presence and intensity. All these studies involved simulations run in a spatially explicit manner, which I believe brought relevant insight on the evolutionary processes investigated. These simulations were coupled to an ABC pipeline, which, here too, proved to be a powerful tool to examine complex evolutionary questions.

Chapter 1 – A simple range-expansion model replicates the general patterns of neutral genetic diversity observed in humans

Ricardo Kanitz, Sylvain Antoniazza, Samuel Neuenschwander, Jérôme Goudet

Manuscript to be submitted to Heredity

A simple range-expansion model replicates complex patterns of neutral genetic diversity observed in humans

Ricardo Kanitz[1,2,]*, Sylvain Antoniazza[1], Samuel Neuenschwander[1,3], Jérôme Goudet[1,2,]*

[1]Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland.

[2]Swiss Institute of Bioinformatics, University of Lausanne, CH-1015 Lausanne, Switzerland.

[3]Vital-IT, Swiss Institute of Bioinformatics, University of Lausanne, CH-1015 Lausanne, Switzerland.

*To whom correspondence may be addressed: ricardo.kanitz@unil.ch, jerome.goudet@unil.ch

Running title: A simple model of human genetic diversity

Keywords: spatially explicit simulations; isolation-by-distance; approximate Bayesian computation; human dispersal; cline vs. clusters.

Abstract

Although it is generally accepted that geography is a major factor shaping human genetic differentiation, it is still disputed whether this differentiation is a result of a simple process of isolation-by-distance, or if there are factors generating distinct clusters of genetic similarity. We address this question using a geographically explicit simulation framework coupled with an Approximate Bayesian Computation approach. Based on six simple statistics only, we estimated the most probable demographic parameters that shaped modern humans evolution under the isolation by distance scenario, and found these were the following: an initial population in East Africa spread and grew from 4000 individuals to 5.7 million in about 132 000 years. Subsequent simulations with these estimates followed by cluster analyses produced results nearly identical to those observed in real data. Thus, a simple diffusion model from East Africa seems to explain a large portion of the genetic diversity patterns observed in modern humans. We argue that a model of isolation by distance along the continental landmasses might be the relevant null model to use when investigating selective effects in humans. From a societal point of view, this model reinforces the idea that there are no different races in our species. Indeed, humans seem to be distributed over a continuum of increasing genetic differentiation.

Introduction

Defining the processes behind the worldwide distribution of human genetic diversity is one of the main ongoing discussions in human population genetics (Handley *et al*, 2007; Rosenberg *et al*, 2005; Serre and Pääbo, 2004). It has long been recognized that geography plays a major role in shaping human genetic diversity (Cavalli-Sforza *et al*, 1994), but it remains unclear whether the patterns observed are sufficiently well explained by isolation-by-distance alone, in a simple process of demographic diffusion of our species, or whether other explanations are needed to understand patterns of human genetic diversity. Indeed, barriers have been put forward as playing an important role in shaping human genetic variation forming more delimited clusters of population structure – i.e. ethnic and continental groups (Rosenberg *et al*, 2005; Rosenberg *et al*, 2002).

This opposition of ideas has generated a debate that has implications for how, if at all, humans are divided in different distinct groups; which in turn has consequences on health policy and research, ethics, and the very understanding of our own species' evolutionary history. Furthermore, studies looking for effects of selection [e.g. (Coop *et al*, 2010; Pickrell *et al*, 2009)] and the association between genotype and phenotype [e.g. (Andersen *et al*, 2012)] strongly depend on models for the underlying neutral evolution. Such models are typically used as null hypotheses and, if incorrect, may lead to erroneous conclusions.

The significance of geography as a shaping agent of human genetic diversity has already been demonstrated in many genetic studies, such as from works based on blood group polymorphism (Cavalli-Sforza and Edwards, 1964), enzyme polymorphism (Nei, 1978), mitochondrial-DNA complete sequences (Ingman *et al*, 2000) up to hundreds of thousands of single-nucleotide polymorphisms (SNP) (Auton *et al*, 2009; Li *et al*, 2008), and even complete genome sequences (The 1000 Genomes Project Consortium, 2010). Nonetheless, which geographical aspects have the largest influence on human genetic diversity is still disputed. Essentially, two competing views could be posed (even though intermediate positions may also be taken): the first one defending the idea in which humans are solely continuously differentiated along a gradient (Handley *et al*, 2007; Prugnolle *et al*, 2005; Serre and Pääbo, 2004); the second, the idea that humans present discrete clusters of genetic differentiation which coincide with continental and sub-continental groupings (Rosenberg *et al*, 2005; Rosenberg *et al*, 2002).

In favor of a clinal view, researchers have based their arguments on the observation that human genetic variability declines as one moves further away from East Africa (Handley *et al*, 2007; Ramachandran *et al*, 2005). Also, it has been observed that there is a clear correlation (R^2=0.85) between genetic distances (e.g., F_{ST}) and geographic distances (along probable colonization routes) (Prugnolle *et al*, 2005). Pro-cluster's arguments do not deny this evidence, but they do see discontinuities along the decline of diversity and argue that major bottleneck events must have generated what one could see as steps in a staircase of genetic diversity (Rosenberg *et al*, 2005). Serre and Pääbo (2004) however brought to discussion the possibility that the geographically uneven sampling scheme seen in most (if not all) worldwide studies on human genetics may have generated false positives for clusters, which would merely reflect the clustered sampling. Rosenberg *et al* (2005) challenged this view taking advantage of an expanded dataset to argue that, among all other variables to be considered in the detection of clusters, geographic dispersion has relatively little effect on the final outcome. In such cases, large amount of genetic data would always allow detecting discontinuities even if the distribution of sampled populations were completely uniform. Such discontinuities could be small, but still detectable and biologically relevant

(Rosenberg *et al*, 2005). Finally, another study more focused on determining the geographical origin of modern humans detected similar patterns of clines in F_{ST} and genetic diversity, and attributed the few deviations from these trends as being caused by "admixture or extreme isolation" (Ramachandran *et al*, 2005).

The lack of agreement on how human neutral genetic diversity is distributed over the globe may bring confusion to studies in related areas. It has been demonstrated that some demographic scenarios might leave signatures which are indistinguishable from those supposedly left by selection. And these demographic scenarios are dependent on the underlying genetic diversity distribution. For instance, Hofer *et al* (2009), looking at four continental human populations, detected an unexpected large proportion of loci (nearly a third) with strong differences in allelic frequency. The authors suggested that the observed patterns are better explained by the combination of demographic and spatial bottlenecks with allele surfing in the front of range expansion rather than by selective factors (Klopfstein *et al*, 2006). In the allele surfing process, drift takes random samples of alleles at potentially different frequencies from the source population (i.e. founder effect), while the combination of range and demographic expansions amplifies this effect on the overall population by increasing the contribution of these alleles in the newly colonized regions. Therefore, to understand the recent evolution of human populations, it is essential to have a good grasp on the neutral events underlying it. A first step to this end is to understand the spatial distribution of human genetic diversity and existence or not of strong discontinuities (i.e. formation of clusters).

Although dense SNP datasets are available, we used a microsatellite dataset in the following investigation for the following reasons: (i) The microsatellites used here have been extensively checked and shown to evolve under the stepwise mutation model

(Pemberton *et al*, 2009), (ii) they are unlinked and essentially neutral, (iii) the number of samples and populations publically available is greater than for SNPs [78 instead of 51 for the latter (Cann *et al*, 2002)], and with better coverage of the American continent, and (iv) we could only simulate so many loci in a spatially-explicit approach with the currently available computational power. Being multi-allelic markers, microsatellites also contain more information per locus than SNPs.

Here, we investigate the distribution of neutral genetic diversity in modern humans using spatially explicit simulations to model the demographic diffusion of our species throughout the globe and to recover the genetic signature left by this process. The simulations are used to estimate, based on six simple and straightforward summary statistics, the demo-genetic parameters best fitting the observed data using Approximate Bayesian Computation (ABC) (Beaumont *et al*, 2002). We do so by generating genetic data under a simple stepping stone model constrained by the shape of the continental masses. Based on the parameter estimates, a second round of simulations is used to generate individual genotypic data ("full dataset"). These data are then subjected to Principal Component Analysis (PCA) and analyses with the STRUCTURE software, where we compared results from the simulations to those obtained for the observation. This allows us to assess the ability of the proposed model to generate the complex patterns observed in the real data. We then discuss the outcomes of such a model for the understanding of the processes defining human genetic diversity around the world and possible applications in the field.

Material and Methods

Observed genetic data. We used 346 microsatellite loci previously verified to evolve according to a stepwise mutation model (Pemberton *et al*, 2009). These loci represent a subset of the data originally made available by Rosenberg *et al* (2002) and Wang *et al*

(2007). The total number of populations in the original dataset was 78, totaling 1484 individuals distributed throughout the world (more details in Figure S1, Figure S2, and Table S1 in the Supplemental Data available online).

ABC. We estimated demographic and genetic parameters using an Approximate Bayesian Computation (ABC) framework. Genetic data were generated using a modified version of quantiNEMO (Neuenschwander *et al*, 2008a) in a two-step process: (i) individual-based forward-in-time simulations for the demography and (ii) coalescent-based backward-in-time simulations to generate the according genetics. Parameters were estimated using the ABC package ABCtoolbox (Wegmann *et al*, 2010).

For the demographic part, all simulations started at one single deme with a varying initial population size (N_i, uniform prior distribution, from 2 to 5120), in Eastern Africa (9°1’48”N, 38°44’24”E) – today’s Ethiopian city of Addis Ababa, the origin of human expansion as estimated by Ray *et al* (2005) and place of the oldest known modern humans remains (Clark *et al*, 2003). The prior distribution for the time for the onset of this expansion had mean 155 000 years ago and standard deviation of 32,000 years (T, generation time of 25 years). These values were based on the combination of independently estimated dates of 141 455 ± 20 000 (Fagundes *et al*, 2007) and 171 500 ± 25 500 years ago (Ingman *et al*, 2000). These dates are more recent than the oldest reliably dated fossil remains in Ethiopia (195 000 ± 5000), which is expected since they most likely predate the spatial expansion of interest in this study (McDougall *et al*, 2005). Dispersal occurs between the four directly neighboring demes in a two-dimensional stepping-stone pattern with a given dispersal rate (m) sampled uniformly between 0 and 0.5. Population regulation followed a stochastic logistic model (Beverton and Holt, 1957) with intrinsic growth rate (r, lognormal prior, mean=0.5, SD=0.6) delimited by the deme’s carrying capacity (N, uniform prior of 2-5120 individuals), used as a proxy for current population size, when multiplied by the total number of habitable demes (5094). For the genetic step, we used a coalescent approach to simulate genealogies for 20 microsatellite loci (single stepwise mutation model with a mutation rate with prior distribution defined by μ (uniform prior of 10^{-5}-10^{-3} mutations/locus/generation) for the same 70 populations and same number of individuals as the observed sampling scheme (details on Table S2).

Summary statistics. In ABC, summary statistics are used to compare observations with simulations (Beaumont, 2010). Ideally, they should be as comprehensive as conceivable in as few values as possible. Initially, we explored a large set of different statistics: number of alleles, allelic richness (Mousadik and Petit, 1996), Garza-Williamson’s M (Garza and Williamson, 2001) and gene diversity (Nei and Chesser, 1983) per sampled population; pairwise F_{ST} (Weir and Cockerham, 1984) and Chord-distances (Nei, 1987) between samples. Considering that many of them did not bring extra information to our model, we retained a subset with the 2415 pairwise F_{ST} between populations and the number of alleles (A) per each one of the 70 demes. These 2485 summary statistics were then transformed into six “pattern” statistics, summarizing the relationships between F_{ST} and pairwise geographic distance. Two linear regressions were made based on these comparisons, from which we then extracted six pattern statistics, namely the means, slopes, and the logarithm of the sum of residuals. The calculations of summary and pattern statistics for the observed data were carried out in the R-package *hierfstat* (Goudet, 2005). Finally, these six pattern statistics were used for the estimates of the demo-genetic parameters and subsequent validations. We also used partial least squares (PLS) to reduce the original 2485 summary statistics to fewer components (Wegmann *et al*, 2009). This technique gave similar (but no better) results for the validations and a few parameters had slightly different estimated values (Figure S4).

Estimates. The six parameters (N_i, μ, m, N, r, T) were estimated based on a comparison of the simulated and the observed summary and a subsequent estimation step. The comparison of the summary statistics was obtained by assessing the Euclidean distance between simulations and the statistics from the observed data, which can be used to rank the simulations from closest to most distant from the observations. Here, we retained the 5000 simulations with smallest Euclidean distances from the observations. This subset of simulations was then used to estimate the parameter values using a weighted generalized linear model (GLM) (Leuenberger and Wegmann, 2010) of the six pattern statistics with the ABCtoolbox software (Wegmann *et al*, 2010).

Validation. We used a validation procedure to assess the quality of our estimates. By using pseudo-observed values taken from the simulations themselves, we verify how well these values could be recovered when estimated through the ABC pipeline used for the real estimates (Neuenschwander *et al*, 2008b). This was done for 1000 different pseudo-observations for each of the six investigated parameters. We calculated then the correlation (R^2) for the regression between pseudo-observed and estimated values, the slope of this regression, the standardized root mean squared error of the mode (SRMSE) and the proportion of estimates for which the 95% higher posterior density interval included the pseudo-observed (“real”) value.

Full-dataset simulations. The estimated parameters were then used to generate a new set of demo-genetic simulations with quantiNEMO, from which we stored the genotypic data at 100 loci for the sampled individuals. We ran three sets of 100 simulations each whose parameter values were sampled from the (i) prior distribution of the estimation step, (ii) posterior distribution (95%HPD) of the estimation step or (iii) taken directly from the point estimates (mode values of the posteriors) of the estimation step. Using the output of these simulations, we ran further analyses in order to compare the simulation's outcome with the patterns observed in the real data. The first comparison of such patterns was based on the six pattern statistics used for the estimations (i.e. mean, slope and sum of residuals for number of alleles and pairwise F_{ST}). We did a second comparison based on the first two axes of a principal component analysis (PCA) computed on the individual allele frequencies in each sampled population. Since the sign of the coordinates along PCA components can differ between replicates, we compared the different sets of simulations by means of the squared correlation between observed and simulated PCA results. Each axis was considered separately. Thus, for each simulation, we estimated an R^2 representing the concordance of simulations and observation in the positioning of the populations on the analyzed PCA axes. These R^2 values were then compared across the three different sets of simulations (Prior, 95%HPD and Mode).

A third comparison was made with population clustering analysis using STRUCTURE v2.3.4 (Pritchard *et al*, 2000). Each simulation was analyzed with the number of clusters K varying from 1 to 7. Each run was made with 250 000 iterations, discarding the first 50 000 as burn-in. The simulations in this comparison where based on the point estimates only. The same strategy was used to analyze the observed data, but for these we used the whole set of 346 microsatellite loci and ran 25 replicates for each K. We post-processed the STRUCTURE outputs with CLUMPP (Jakobsson and Rosenberg, 2007) in order to align the different replicates and also summarize them into one final output which was then used to compare simulations with the observations. We also carried out the estimation of the number of groups (K) best explaining the variation present in simulations and observations following Evanno *et al* (2005). The ΔK was estimated based on 25 replicates for each STRUCTURE run.

Results

Parameter estimates and validation. We ran in total 1 183 831 simulations based on prior distributions; 974 934 (82.4%) successfully colonized all the sampled patches and were therefore used in the subsequent analyses. We obtained posterior estimates for all six demo-genetic parameters, which are presented in Table 1 (point estimates; for their complete distributions, see Figure S3). Briefly, the time of expansion T is estimated to be 132 250 years before present, the initial population size N_i close to 4000 individuals, the current world population effective size N slightly more than 5.7 million individuals; the mutation rate μ is estimated at $2.6x10^{-4}$, the population growth rate r at 0.149 and the migration rate between neighboring populations m at 0.041. To assess the quality of these parameter estimates, we performed a validation step (Wegmann *et al*, 2010). This assessment, based on 1000 independent simulations, allowed grouping the results into three qualitative groups. Excellent estimability was attained for mutation rate (μ) since we observed a strong correlation between pseudo-observations and estimations (R^2=0.877) for which the slope was nearly 1 (slope=0.908), the error rate was low (SRMSE=0.099), and the proportion of the estimates that included the pseudo-observed value within their 95%HPD interval was 0.977, suggesting only slightly more conservative posteriors. Good estimability was also achieved with migration rate (m), current population size (N) and initial population size (N_i) for which the R^2 values were about 0.5 and the slopes above 0.6. We had rather poor estimability for time of the onset (T) and population growth rate (r) where R^2 values were below 0.3 (Table 1).

Full-dataset simulations. The posterior estimates above were then used in further simulations producing complete genotypes (not only summary statistics) for all sampled individuals at 100 simulated microsatellite loci. These additional simulations were carried-out by randomly sampling parameter values from the prior and truncated posterior (at the 95%HPD level) distributions and also by directly using the point estimates. We first addressed whether our simulations could replicate the patterns already observed in the original genetic data. Here, we extended the analyses of the observed dataset to a more complete coverage than previous studies [as reviewed in Handley *et al* (2007)], with the addition of 21 Native American populations from Wang *et al* (2007). Figure 1A shows the observed patterns of reduction of genetic diversity and isolation by distance, while Figure 1B shows the comparison with a typical full-dataset simulation rerun with parameter values based on the point estimates of the parameters. For both observation and simulations, the general pattern is the same: a steady reduction of diversity for populations as one moves away from Addis Ababa, and a clear-cut increase of genetic differentiation with geographic distance. The comparison between these simulations' results served as proxy for the convergence of the parameter estimates: As expected, we observe that, with more restrictive samplings of parameter values (from sampling in the prior distribution to sampling in the posterior distribution to using the point estimate), the statistics in the simulations better approximate the values of the observed data (Figure 2 and Figure S5).

Next, we investigated whether the simulated genetic data could reproduce the patterns observed in a Principal Component Analysis (PCA) of the observed data set. In the observed dataset, one observes clear divisions between continental groups (Figure 3A), as previously demonstrated elsewhere (Biswas *et al*, 2009; Li *et al*, 2008). The PCA results based on our simulations returned a pattern very similar to that observed (Figure 3A). The convergence (from prior to 95% HPD to point estimate) of parameter estimates can also be assessed with PCA: The correlation between observation and simulations in their principal components (PC1 and PC2) are presented in Figure 3B. For the first component, the correlation was similar for the three groups of simulations (parameters drawn from the prior, the posterior or the mode of the posterior distribution); for the second, there was a trend of higher correlation as simulations based on more restrictive samples of the posterior distribution were used.

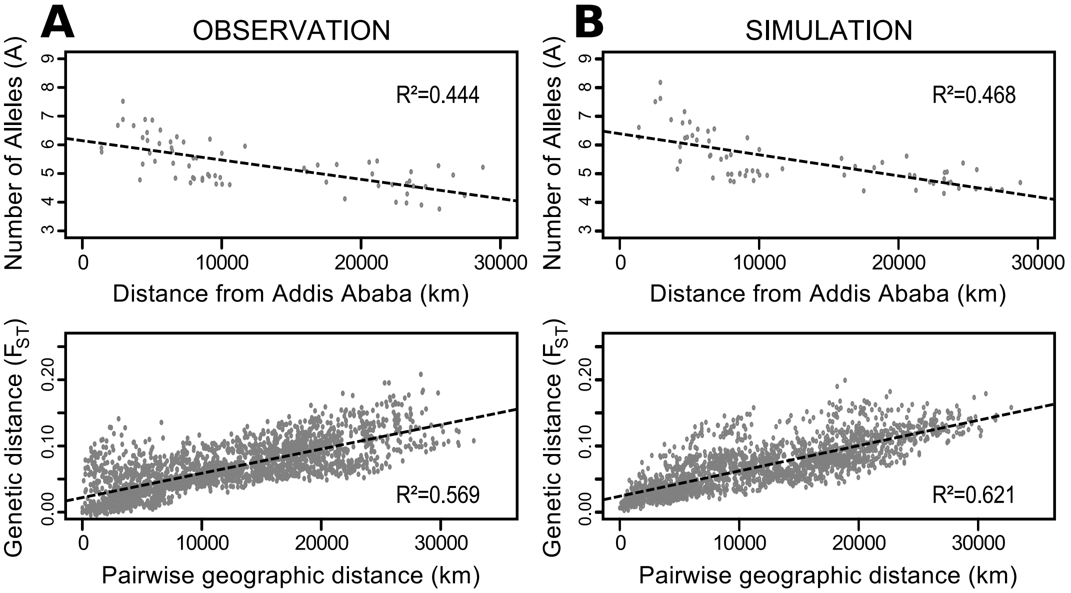

Figure 1: Comparison of the patterns of isolation by distance generated with the observed and simulated data. In **A**, the patterns obtained for the observed data; in **B**, the result of one of the simulations based on the point estimates. Each point represents a population (top) or a pairwise population comparison (bottom); the dashed lines represent the linear regressions of these points (whose R^2 values are informed).

Finally, we also looked at the partitioning pattern generated by the software STRUCTURE. Simulations and observation gave the same estimates of the most likely number of groups (K) within the worldwide sample either using the highest likelihood of the data as the criteria for defining K (which led to K=7 in both observations and simulations); or using ΔK (Evanno *et al*, 2005), which favored K=2 both for observations and simulations (Figure S7). The similarities also persist in the way the different individual genomes are allocated to the different clusters resulting from this analysis. They generated, for both observed and simulated data, remarkably similar results for K=2 to K=4 (Figure 4). For K=2, African and Americans individuals have genomes entirely assigned to one of the clusters, whereas all other individuals are admixed to different extents, and the proportion of admixture in the simulations matches almost perfectly that in the observation. For K=3, Eurasian populations emerge from the other groups previously formed with a few differences between simulations and observation: In the observations, Middle-Easterners and Europeans group with Africans; whereas in the simulations, they are admixed between the African and East Asian clusters. For K=4, the African component becomes clear by dividing the sub-Saharan samples from the rest of the world. Whereas for the observation this division is very clear, the results based on the simulated data show a more gradual pattern with Middle-Eastern and European mixed-ancestry samples. Beyond K=4, the patterns observed between simulations and observations diverge: while single populations start to emerge as separate groups in the observation; higher values of K generated admixed individuals and populations within the already existing groups in the simulations (Figure S6). Interestingly, in both simulations and observation, the grouping pattern is relatively consistent with the continental partitioning of the populations.

Discussion

We have shown that a simple diffusion model along landmasses generates genetic patterns very similar to those observed in the real dataset. The signatures of isolation-by-distance and constant decrease of genetic diversity with increasing distances from Addis Ababa retrieved from the simulations show remarkable similitude with the observations. Importantly, these similarities are not restricted to the statistics used to estimate the demo-genetic parameters, but are also present in the other analyses we ran. The PCA results for the simulations based on the modes of the posterior distributions show a strong correlation with both the first and second principal components calculated from the observation. And the analyses using the software STRUCTURE presented highly similar results for observation and simulations: they are consistent in the number of groups which better explains the diversity in the samples and also show very similar population divisions up to four clusters.

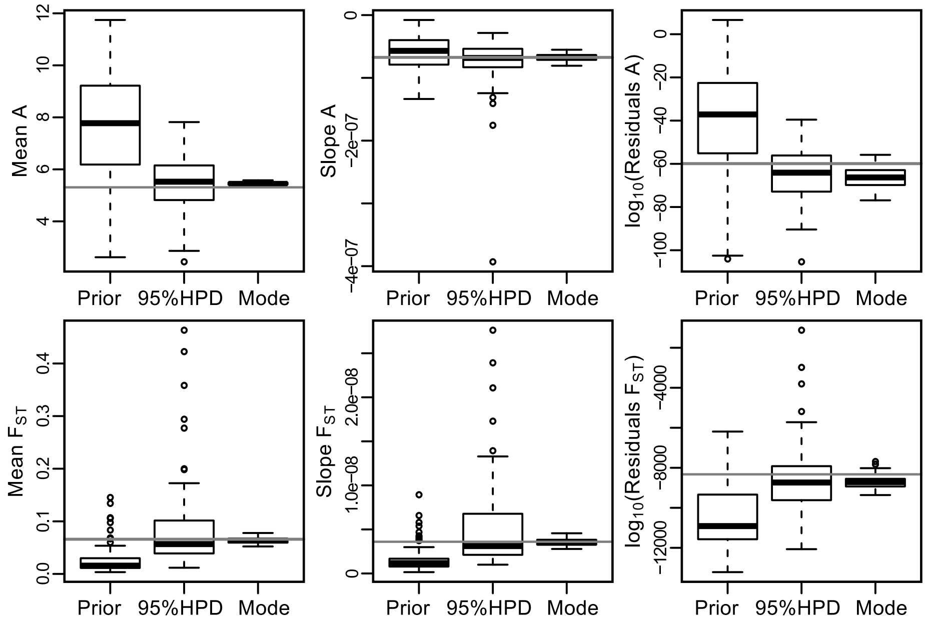

Figure 2: Convergence of the pattern statistics towards the observation (horizontal gray line) in different sets of simulations. Within each plot, we present the different sources for the simulations that generated the distributions: “Prior” are simulations sampled randomly from the whole prior; “95%HPD” are simulations run based on the 95% higher posterior density estimates for all parameters; and “Mode” are simulations based on the point estimates for all parameters.

Patterns from the full-dataset simulations. PCA has long been used in human population genetics (Menozzi *et al*, 1978). Even though the interpretations made on those first results are questionable when it comes to detecting migration events (Novembre and Stephens, 2008), it is clear that PCA is able to relate genetic variation to the geographic distribution of populations (Novembre *et al*, 2008) and even individuals (Wang *et al*, 2012). Here, we compared the positioning of the sampled populations on the first two principal components between simulations and observation. Their resemblance seen in Figure 3A is clear and, even though it is based on one of many simulations, it is by no means an atypical example; to the contrary, it is one of many that are very similar to the observation. Over all sets of simulations, the coordinates of the samples along the first component (Figure 3B) show a very high correlation with the observed coordinates, even for simulations based on the prior, uninformative, distribution of the parameters. This indicates that the first axis of the PCA (capturing the largest fraction of the genetic variance) probably relates to the origin of the expansion (which occurs in the same place, East Africa, for all simulations) and demic diffusion. The second principal component seems to be more sensitive to the choice of the parameter values, the correlation between observation and simulations increasing when the parameters used for the simulations get closer to the estimation.

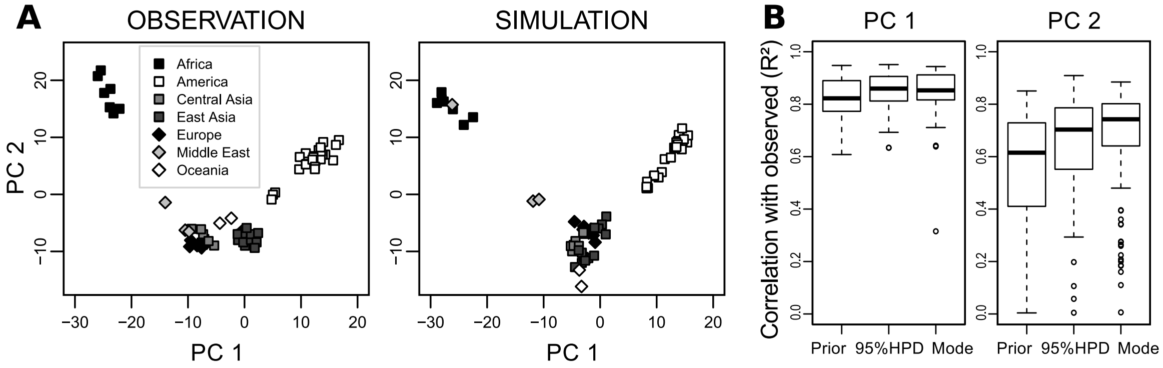

Figure 3: PCA results in observation and simulations. **A**, Comparison of PCA applied to the observed data (left) and one selected simulation (right). The first (PC 1) and second (PC 2) principal components are represented here, where each point represents one of the analyzed populations, grouped by continents. **B**, Boxplots of the correlation values between the two first principal components in observations and simulations based on the prior distribution (“Prior”), 95% higher posterior density distribution (“95%HPD”), and on the point estimates (“Mode”).

Even having in mind that admixture-based analyses are not completely independent from PCA (Engelhardt and Stephens, 2010), the most surprising result obtained here comes from the population clustering analysis in STRUCTURE. We were uncertain of the possibility of clusters to appear as a result of a simple diffusion process such as that we used in our simulations. In fact, based on ΔK, the estimation of the best number of groups is K=2 (possibly K=1), which suggests the inexistence of separate genetic groups in the simulations. Importantly, this is also the case for the real dataset, which has been – regardless of that – consistently analyzed as if there were more genetic clusters in it (Rosenberg *et al*, 2005). We therefore also compared the simulations with the observations with higher values of K and found a high consistency in the order in which new clusters appear. The American populations are the first to stand out; second, a separation between European and African versus East Asian; and then the Africans alone stand out from the rest. There are a few exceptions though. The Mozabite population, from North Africa, tends to group with the other African populations in the PCA results for the simulations; while, in the observed data, they group with the Middle-Eastern and European populations. It is possible that more recent events of contact through the Strait of Gibraltar (Currat *et al*, 2010) or the Fertile Crescent, which are not captured by our simulations, contributed to this discrepancy. Another explanation could be the absence of the potentially important barrier of the Sahara desert in the simulations, which, in reality may have played an important role in isolating North Africans from sub-Saharan populations. Studies more focused in this region should take these possibilities into account when looking at the local populations' genetic composition. Besides that, the European/Middle-Eastern samples present a more mixed composition in the simulations. Here, again, the presence/absence of the Sahara might have its importance, and we know from other sources that the history of the peopling of Europe, the fertile crescent and North Africa is more complex (Arenas *et al*, 2012). But the key point remains that, on the whole, simulations and observed data lead to the same splits among human groups.

Table 1: Accuracy table and point estimates for the six variable parameters in the ABC framework. Point estimate corresponds to the mode of the posterior distribution, while HPD95% interval represents the parameter values comprised within the 95% higher posterior density interval. R^2 stands for the coefficient of determination of pseudo-observed on estimated values; SRMSE is the root mean squared error of the mode, standardized between 0 and 1; Prop. HPD95% stands for the proportion of tests for which 95% higher posterior density intervals include the true value. All rates are per generation (25 years).

	T (years)	**N_i (ind.)**	**N (ind.)**	**μ**	**r**	**m**
Point estimate	**132 250**	**3952**	**5 725 656**	**2.6×10^{-4}**	**0.149**	**0.041**
HPD95% interval	60 850 - 203 900	920 - 5120	35 658 - 20 905 776	9.3×10^{-5} - 4.4×10^{-4}	0.036 - 0.679	0 - 0.177
R^2	0.235	0.399	0.431	0.877	0.286	0.57
SRMSE	0.132	0.233	0.227	0.099	0.108	0.187
Slope	0.248	0.536	0.602	0.908	0.352	0.682
Prop. HPD95%	0.993	0.956	0.981	0.977	0.983	0.979

It is essential to mention that when looking at the most probable number of groups according to the Delta K method (K=2) the results obtained for simulation and observation are virtually indistinguishable. Meaning that, for the most relevant part of the overall genetic differentiation for both simulations and observation (i.e. two groups), the resulting assignment of the populations to one or the other group is fundamentally the same.

As mentioned above, to use microsatellite loci in this study was an informed decision taken on the basis of the effort to improve the amount of information captured with the limited number of loci that could be simulated. A comparison of results obtained in previous studies across these two kinds of markers is nonetheless possible. For the PCA results, studies on SNP worldwide datasets (Biswas *et al,* 2009; Jakobsson *et al*, 2008; Li *et al*, 2008; Wang *et al*, 2012) return results very similar to the results obtained here both for the observed and simulated data (Figure 3A). For the admixture-based STRUCTURE analyses, the similarities across markers remain: Rosenberg *et al*

(2005) using microsatellite data have found results very similar to those obtained with SNPs in Li *et al* (2008), which are, in turn, very similar to our results in Figure 4. As it seems, for capturing the overall human genetic distribution, the SNP data may increase the resolution of the results (Li *et al*, 2008), but does not seem to affect the general patterns that are replicated in the model we propose here.

Selection. Detecting and measuring natural selection in action in the human genome has been one of the main goals of population geneticists in the last decades (Nielsen *et al*, 2007; Sabeti *et al*, 2007; Voight *et al*, 2006). Some of the methods for detecting selection are based on the comparison of F_{ST} across loci [e.g. (Foll and Gaggiotti, 2008)]. Observed markers are compared against a neutral reference (or null) distribution of F_{ST} values simulated using a simple island model. Loci whose F_{ST} values are too low or too high are then considered to be, respectively, under balancing or directional selection. The power of this method is negatively affected with increasing deviation of the real demography from the simulated island model. Excoffier *et al* (2009), for instance, showed that the addition of one layer of complexity to the base model, making it hierarchical with two levels, already reduces the number of false-positives to very large extent. Possibly, at least for humans, a better underlying model would be a clinal one. Including null model with isolation by distance seems the next key step to improve methods for the detection of selection.

While it seems clear that additional spatial heterogeneity could help improving a basic neutral model (by accounting, e.g., for the Sahara), the model we used here proved to be a very useful one for explaining many patterns of human genetic variation. Such a model may represent a good choice for establishing a neutral background in future studies looking at more complex questions in modern human evolution such as the detection of selective events.

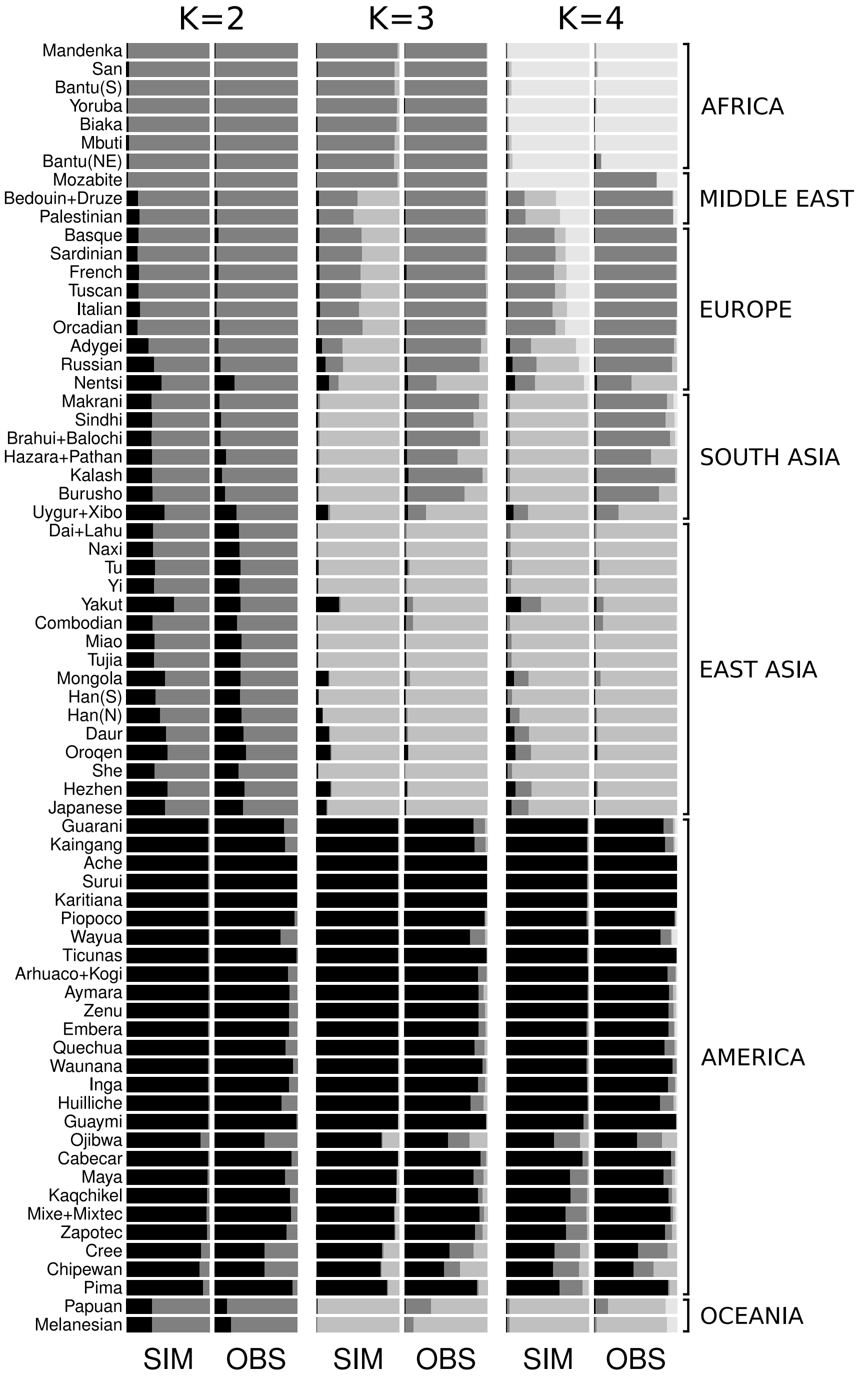

Figure 4: Comparison between the STRUCTURE results obtained for observed (OBS) and simulated (SIM) data. Horizontal bars represent the 70 populations as used in the simulations and the different shades of gray code for the proportion of each inferred ancestry group (K from 2 to 4).

Cline vs. clusters. The results obtained here shed new light on the “cline vs. clusters” controversy. The fact that a simple model of two-dimensional dispersion on a homogeneous world succeeds in producing results so similar to the real data in many different analyses is strong support for an overall clinal view of the distribution of human genetic diversity over the globe. Even though the simulations used here involve some sophistication, the model they suggest is simple and can easily be considered in further population genetics studies: isolation-by-distance and continuous decline of diversity as we move away from East Africa. These two patterns are easily described by two linear regressions after all.

The clinal model for the global distribution of human diversity encounters support in other biological and cultural systems. Skull morphological diversity, for example, shows a clear and steady decline of within population diversity as the distance from Africa increases and is in perfect agreement with what is found on DNA (Betti *et al*, 2009). Language, a cultural feature, also shows a similar pattern. Distance from Africa, alone, explains 30% of the reduction in phonemic diversity as measured in 504 languages worldwide (Atkinson, 2011).

This view of human genetic diversity distributed over a continuous cline reinforces the notion of inexistence of biological races (Group, 2005; Jorde and Wooding, 2004; Long and Kittles, 2009). Nonetheless, classifying humans in different groups is still common practice in many genetic studies (mostly medical genetics). As most of these studies are actually dealing with local populations sampled in discontinuous ways or with immigrants whose origins are very different around the world, assuming discrete groups is not necessarily an incorrect approach. However, if one is interested in moving to broader scale studies, putting the subjects in different “boxes” may lead to mistakes and misconceptions.

Supplemental data

Supplemental Data include seven figures and two tables and can be found further bellow in this chapter.

Acknowledgements

We would like to thank Nelson J. R. Fagundes for commenting on previous versions of this manuscript and to John Pannell for fruitful discussions. Also, we are grateful to Anna Kostikova who independently verified the calculation of geographic distances between populations. Funding was provided by the Swiss National Science Foundation grants N° 31003A-138180 and 31003A-108194 to JG. The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

References

Andersen KG, Shylakhter I, Tabrizi S, Grossman SR, Happi CT, Sabeti PC (2012). Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **367**(1590)**:** 868-877.

Arenas M, François O, Currat M, Ray N, Excoffier L (2012). Influence of admixture and Paleolithic range contractions on current European diversity gradients. *Molecular Biology and Evolution*.

Atkinson QD (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* **332**(6027)**:** 346-349.

Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A *et al* (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res* **19**(5)**:** 795-803.

Beaumont MA (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics* **41**(1)**:** 379-406.

Beaumont MA, Zhang W, Balding DJ (2002). Approximate Bayesian Computation in Population Genetics. *Genetics* **162**(4)**:** 2025-2035.

Betti L, Balloux F, Amos W, Hanihara T, Manica A (2009). Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proceedings Biological sciences / The Royal Society* **276**(1658)**:** 809-814.

Beverton RJH, Holt SJ. (1957). *Fisheries Investment Series, Vol. 19*. U.K. Ministry of Agriculture and Fisherie: London.

Biswas S, Scheinfeldt LB, Akey JM (2009). Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *American journal of human genetics* **84**(5)**:** 641-650.

Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L *et al* (2002). A Human Genome Diversity Cell Line Panel. *Science* **296**(5566)**:** 261-262.

Cavalli-Sforza LL, Edwards AWF (1964). Analysis of human evolution. *Proc XI Internat Congr Genet* **3:** 923-933.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). *The History and Geography of Human Genes*. Princeton University Press.

Clark JD, Beyene Y, WoldeGabriel G, Hart WK, Renne PR, Gilbert H *et al* (2003). Stratigraphic, chronological and behavioural contexts of Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423**(6941)**:** 747-752.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**(4)**:** 1411-1423.

Currat M, Poloni E, Sanchez-Mazas A (2010). Human genetic differentiation across the Strait of Gibraltar. *BMC evolutionary biology* **10**(1)**:** 237.

Engelhardt BE, Stephens M (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* **6**(9)**:** e1001117.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**(8)**:** 2611-2620.

Excoffier L, Hofer T, Foll M (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* **103**(4)**:** 285-298.

Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL *et al* (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**(45)**:** 17614-17619.

Foll M, Gaggiotti O (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* **180**(2)**:** 977-993.

Garza JC, Williamson EG (2001). Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology* **10**(2)**:** 305-318.

Goudet J (2005). hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**(1)**:** 184-186.

Group REaGW (2005). The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research. *American journal of human genetics* **77**(4)**:** 519-532.

Handley LJ, Manica A, Goudet J, Balloux F (2007). Going the distance: human population genetics in a clinal world. *Trends in genetics : TIG* **23**(9)**:** 432-439.

Hofer T, Ray N, Wegmann D, Excoffier L (2009). Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet* **73**(1)**:** 95-108.

Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* **408**(6813)**:** 708-712.

Jakobsson M, Rosenberg NA (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**(14)**:** 1801-1806.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC *et al* (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**(7181)**:** 998-1003.

Jorde LB, Wooding SP (2004). Genetic variation, classification and 'race'. *Nature genetics*.

Klopfstein S, Currat M, Excoffier L (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution* **23**(3)**:** 482-490.

Leuenberger C, Wegmann D (2010). Bayesian computation and model selection without likelihoods. *Genetics* **184**(1)**:** 243-252.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S *et al* (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**(5866)**:** 1100-1104.

Long JC, Kittles RA (2009). Human genetic diversity and the nonexistence of biological races. *Hum Biol* **81**(5-6)**:** 777-798.

McDougall I, Brown FH, Fleagle JG (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433**(7027)**:** 733-736.

Menozzi P, Piazza A, Cavalli-Sforza L (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201**(4358)**:** 786-792.

Mousadik A, Petit RJ (1996). High level of genetic differentiation for allelic richness among populations of the argan tree [Argania spinosa (L.) Skeels] endemic to Morocco. *Theoret Appl Genetics* **92**(7)**:** 832-839.

Nei M (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**(3)**:** 583-590.

Nei M (1987). *Molecular Evolutionary Genetics*. Columbia University Press: New York.

Nei M, Chesser RK (1983). Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**(3)**:** 253-259.

Neuenschwander S, Hospital F, Guillaume F, Goudet J (2008a). quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. *Bioinformatics* **24**(13)**:** 1552-1553.

Neuenschwander S, Largiader CR, Ray N, Currat M, Vonlanthen P, Excoffier L (2008b). Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. *Mol Ecol* **17**(3)**:** 757-772.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007). Recent and ongoing selection in the human genome. *Nat Rev Genet* **8**(11)**:** 857-868.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A *et al* (2008). Genes mirror geography within Europe. *Nature* **456**(7218)**:** 98-101.

Novembre J, Stephens M (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* **40**(5)**:** 646-649.

Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg NA (2009). Sequence determinants of human microsatellite variability. *BMC genomics* **10:** 612.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D *et al* (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* **19**(5)**:** 826-837.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**(2)**:** 945-959.

Prugnolle F, Manica A, Balloux F (2005). Geography predicts neutral genetic diversity of human populations. *Current biology : CB* **15**(5)**:** R159-160.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **102**(44)**:** 15942-15947.

Ray N, Currat M, Berthier P, Excoffier L (2005). Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res* **15**(8)**:** 1161-1167.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* **1**(6)**:** e70.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA *et al* (2002). Genetic structure of human populations. *Science* **298**(5602)**:** 2381-2385.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C *et al* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**(7164)**:** 913-918.

Serre D, Pääbo S (2004). Evidence for Gradients of Human Genetic Diversity Within and Among Continents. *Genome Research* **14**(9)**:** 1679-1685.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**(7319)**:** 1061-1073.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006). A map of recent positive selection in the human genome. *PLoS Biol* **4**(3)**:** e72.

Wang C, Zöllner S, Rosenberg NA (2012). A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS Genetics* **8**(8)**:** e1002886.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G *et al* (2007). Genetic variation and population structure in native Americans. *PLoS Genet* **3**(11)**:** e185.

Wegmann D, Leuenberger C, Excoffier L (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**(4)**:** 1207-1218.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC bioinformatics* **11:** 116.

Weir BS, Cockerham CC (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**(6)**:** 1358-1370.

Supporting Information

A simple range-expansion model replicates complex patterns of neutral genetic diversity observed in humans

Ricardo Kanitz, Sylvain Antoniazza, Samuel Neuenschwander, Jérôme Goudet

Supplementary Figures

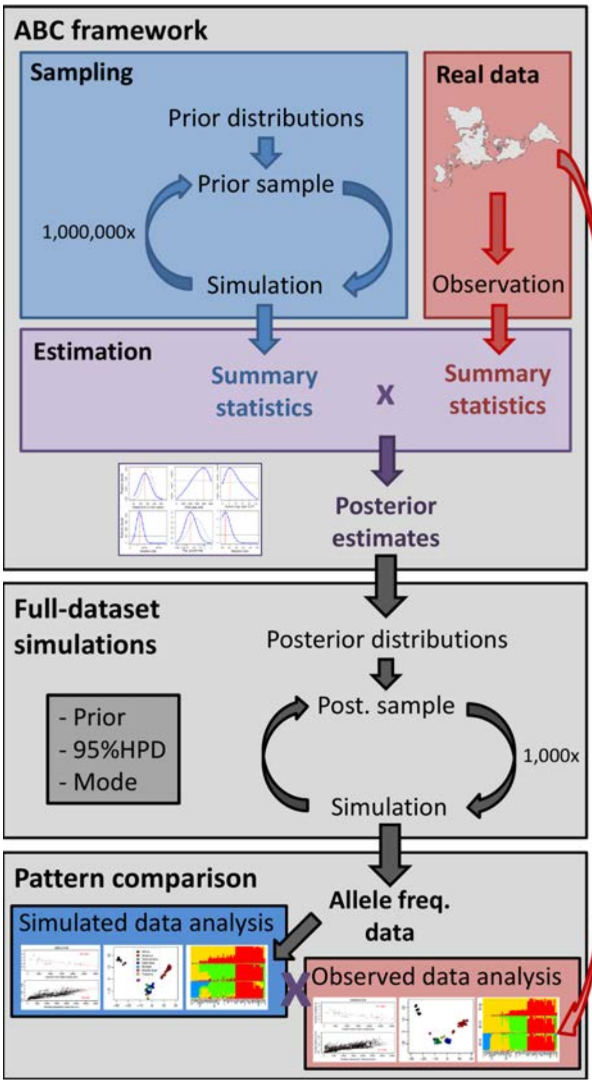

Figure S1: Schematic representation of the pipeline used in the study. **ABC framework** shows the basic structure of an ABC analysis focused in parameter estimation. **Full-dataset simulations** represents the following step in which simulations were run based on the estimations above and for which complete allele frequency data was retained. In **Pattern comparison**, further analyses were run in order to compare simulations and observations in way they produce results for IBD regression analysis, PCA and STRUCTURE.

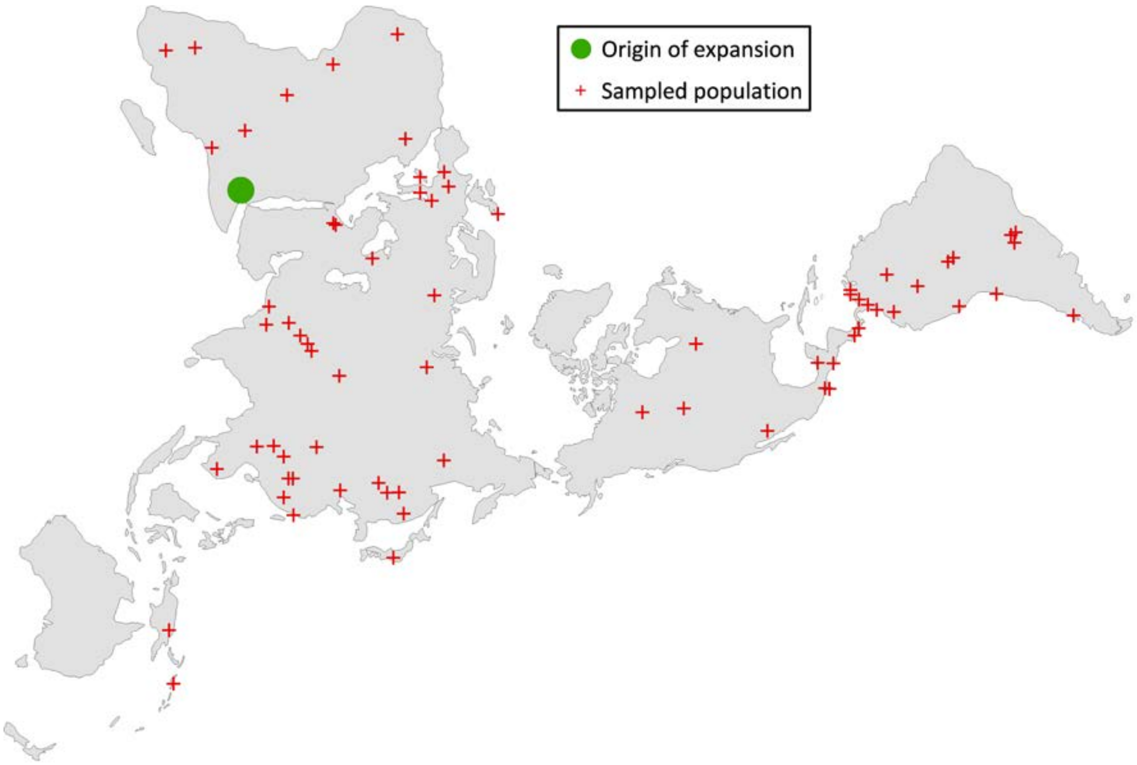

Figure S2: Distribution of the populations used in this study (red crosses). The origin of the expansion of humans in East Africa is marked as the green dot. Map following Fuller's Dymaxion projection, the same applied to the maps used in the simulations. The modeled map contained 20,384 square demes (5,094 on land), each with an approximate area of 160 x 160 km^2. The pairwise geographic distances between populations were calculated with the R package *gdistance* correcting for the Earth curvature and considering only on-land pathways – and between A and geographic distance from Addis Ababa (the origin of the expansion).

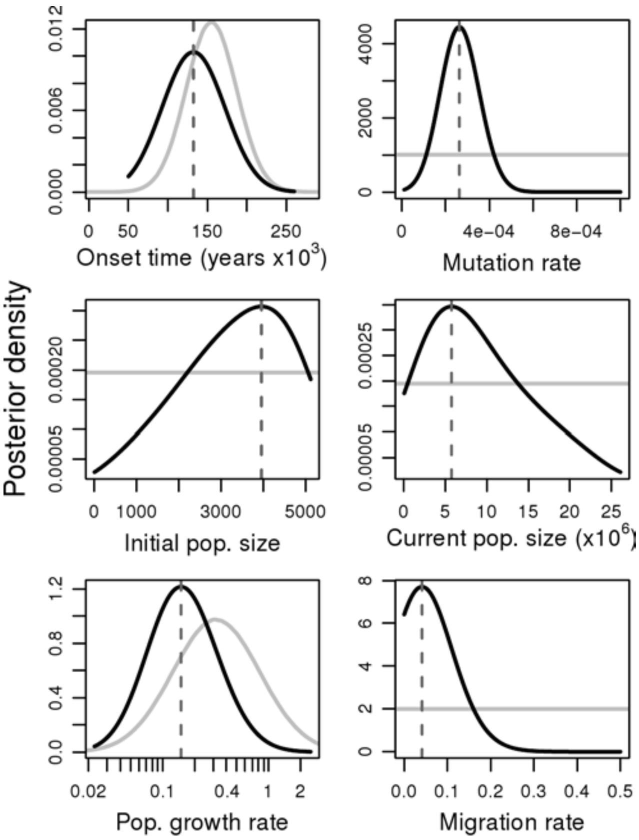

Figure S3: ABC-GLM estimation of the model parameters. Gray lines represent the prior distributions; black lines, the posteriors; the gray dashed vertical lines, the modes for the posteriors (point estimates). The estimations were carried out on 5,000 out of ~1 million simulations which were the closest to the observations in six pattern statistics (see material and methods for details).

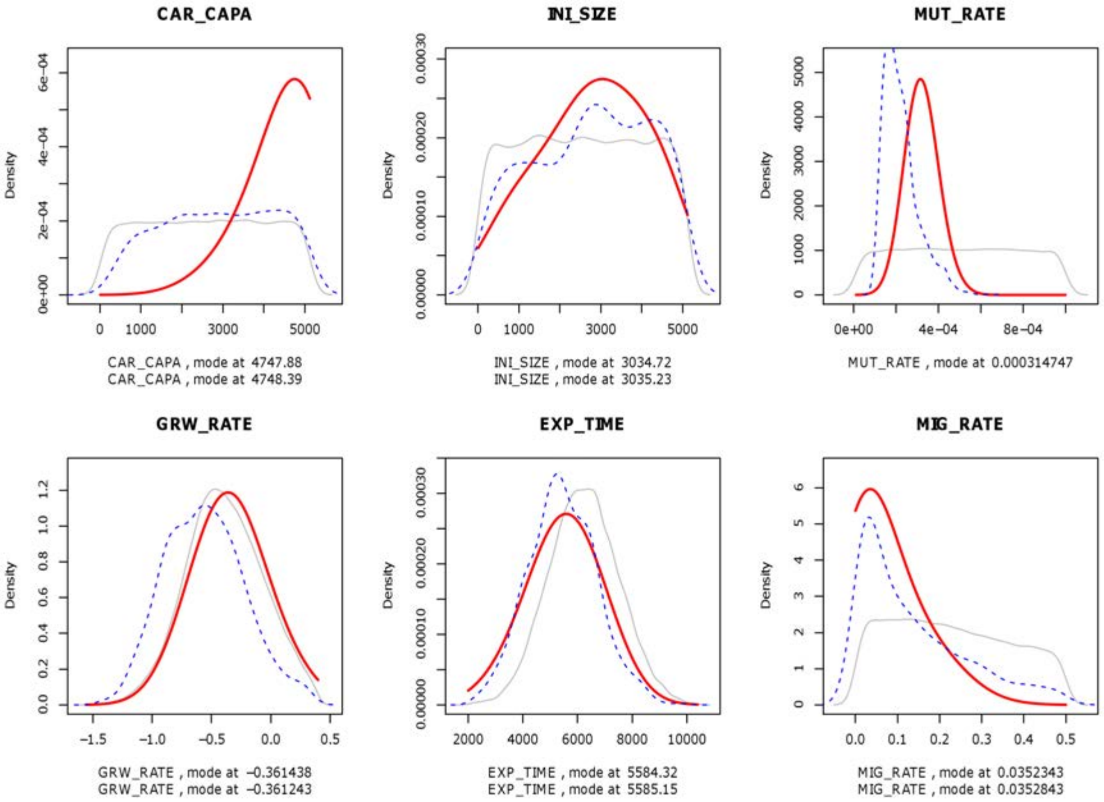

Figure S4: ABC-GLM estimation of the model parameters using five PLS components calculated from the whole set of statistics retaining 1000 simulations. Gray lines represent the realized priors; blue dashed lines represent the distribution of the parameter values in the retained simulations; red lines represent the posterior distributions. The PLS calculation was conducted on a set of 2,485 statistics comprising number of alleles (A) and gene diversity (Hs) per patch and all pairwise F_{ST} comparisons between patches. **CAR_CAPA** stands for current population size; **INI_SIZE**, initial population size; **MUT_RATE**, mutation rate; **GRW_RATE**, population growth rate; **EXP_TIME**, time of onset of the expansion; **MIG_RATE**, migration rate. Below each panel, the values for the mode (point estimates) are given for every parameter.

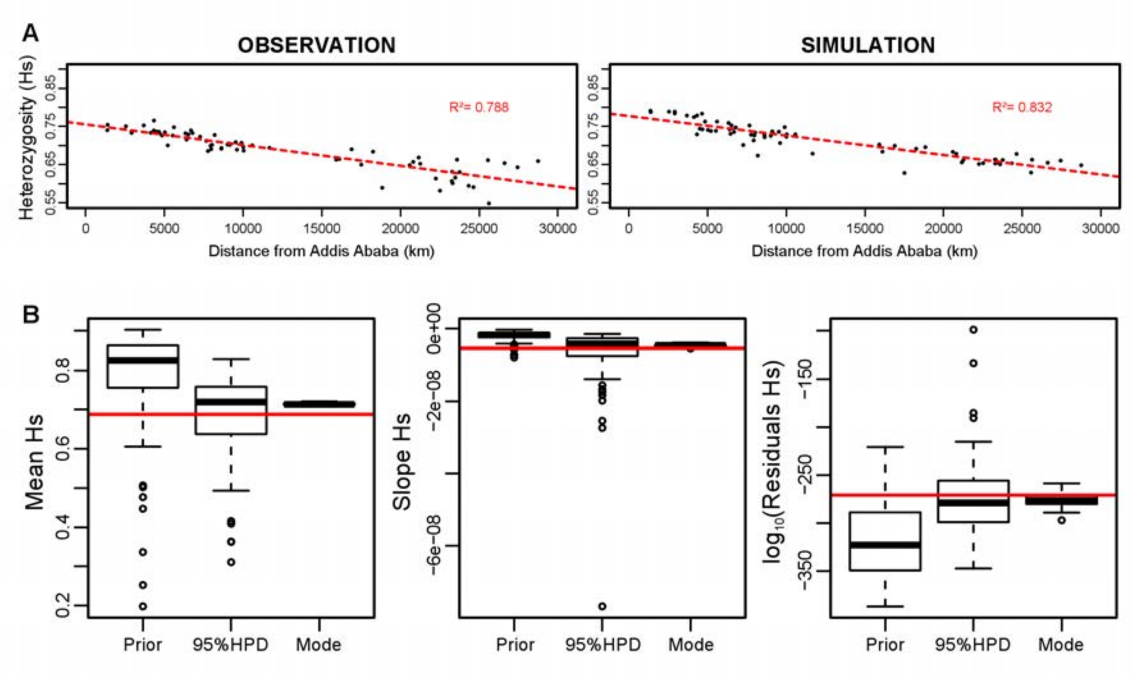

Figure S5: Comparison of patterns generated with gene diversity (heterozygosity, hs). **A**, comparison of the patterns generated for the cline in heterozigosity between observation and a simulation based on the point estimates. **B**, convergence of different pattern statistics related to the heterozigosity cline across different samplings from prior or posterior.

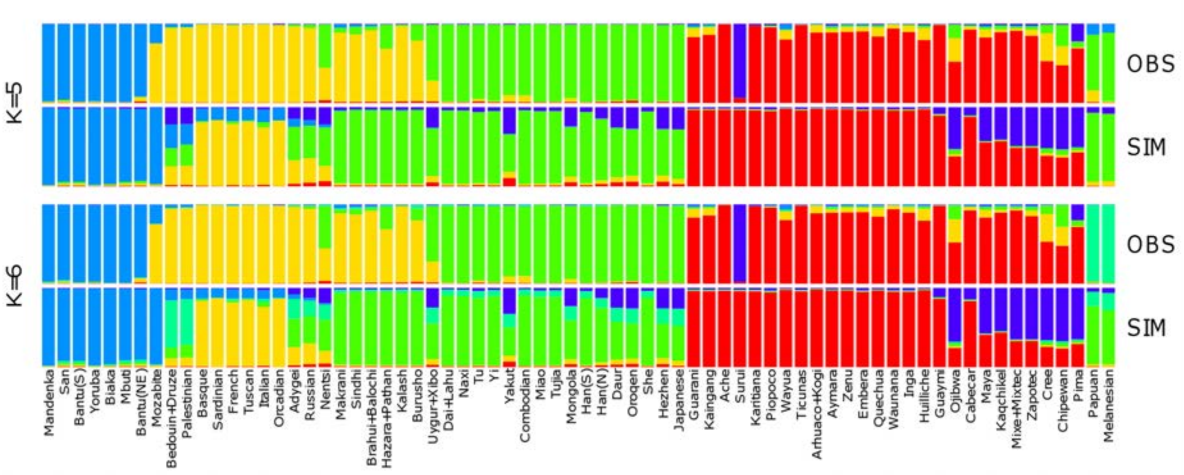

Figure S6: Comparison between the STRUCTURE results obtained for observed (OBS) and simulated (SIM) data. Vertical bars represent the 70 populations as used in the simulations and the colors code for the proportion of each inferred ancestry group (K = 5 and 6). One can observe that particular populations become highlighted in the observations (Suruí with K=5, Oceanians with K=6); while, in the simulations, many populations begin to show admixed compositions.

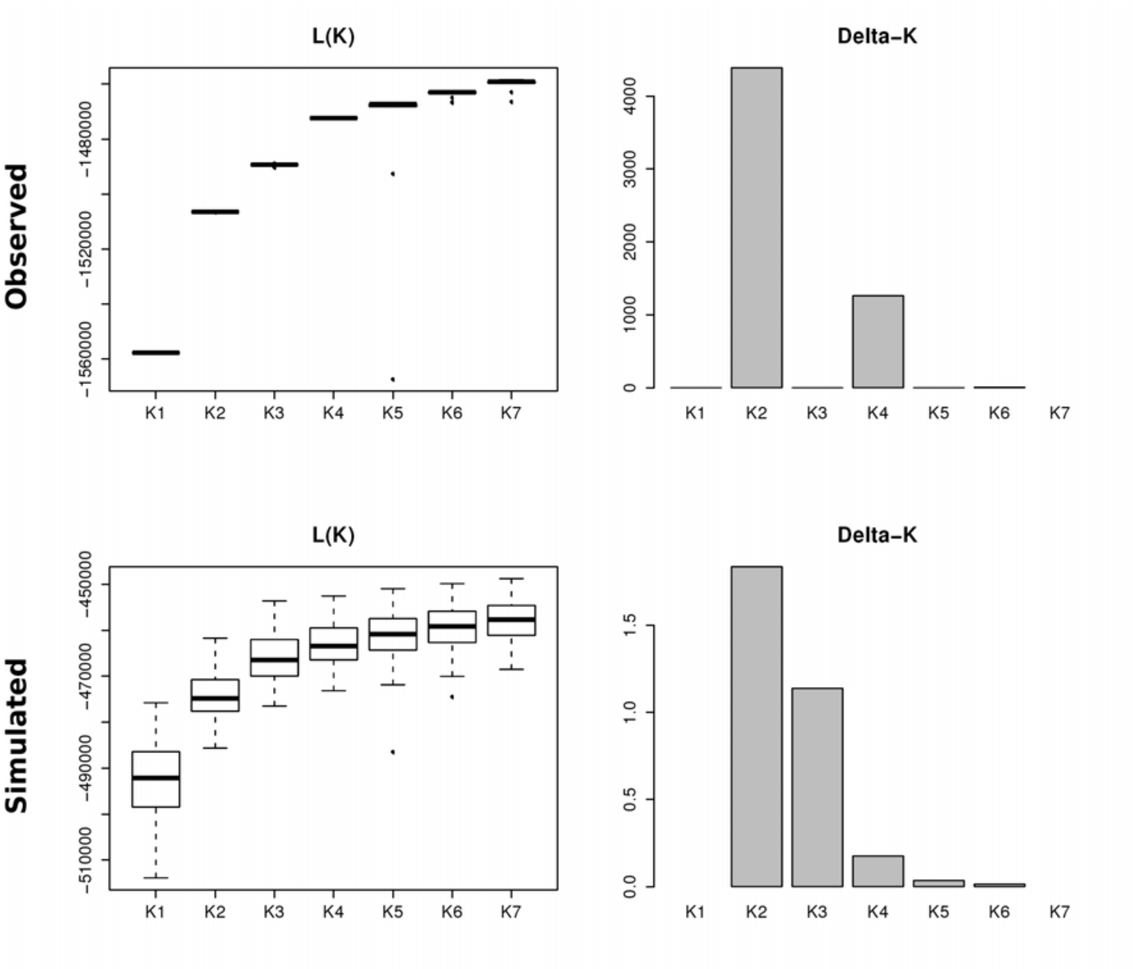

Figure S7: Estimates of the most likely number of groups within the worldwide sample of populations. The figure contains the results obtained both for observations (Observed) and simulations (Simulated). L(K) is the direct assessment of likelihood for each number of groups. Delta-K is the estimate based on Evanno et al.’s 2005 approach.

Supplementary Tables

Table S1: Population samples as they were analyzed in this study. Populations marked with “a” were merged together due to their geographical proximity (less than 160km apart) and were considered to inhabit the same deme in the simulations and also in the analyses applied to the read dataset. Populations marked with “b” were removed from the pattern statistics calculations: They were either known exceptions to the general patterns found in the continent (Aché), or were sampled in the vicinity of other populations, on the edges of their original distributions. For these, we kept the populations with the larger sample sizes and these were the Karitiana (as opposed to the Suruí) and Guarani (as opposed to the Kaingang).

Continent	Population	Number of individuals
Africa	Bantu North-Eastern Africa	12
Africa	Bantu Southern Africa	8
Africa	Biaka Pygmies	32
Africa	Mandenka	24
Africa	Mbuti Pygmies	15
Africa	San	7
Africa	Yoruba	25
America	Ache[b]	19
America	Arhuaco & Kogi[a]	34
America	Aymara	18
America	Cabecar	20
America	Chipewan	29
America	Cree	18
America	Embera	11
America	Guarani	10
America	Guaymi	18
America	Huilliche	20
America	Inga	17
America	Kaingang[b]	7
America	Kaqchikel	12
America	Karitiana	24
America	Maya	25
America	Mixe & Mixtec[a]	40
America	Ojibwa	20
America	Pima	25
America	Piopoco	13
America	Quechua	20
America	Surui[b]	21

America	Ticuna-Arara & Ticuna-Tarapaca[a]	35
America	Waunana	20
America	Wayua	17
America	Zapotec	19
America	Zenu	18
South-Asia	Brahui & Balochi[a]	50
South-Asia	Burusho	25
South-Asia	Hazara & Pathan[a]	48
South-Asia	Kalash	25
South-Asia	Makrani	25
South-Asia	Sindhi	25
South-Asia	Uygur & Xibo[a]	19
East-Asia	Combodian	11
East-Asia	Dai & Lahu[a]	20
East-Asia	Daur	10
East-Asia	Han Central China	34
East-Asia	Han Northern China	10
East-Asia	Hezhen	9
East-Asia	Japanese	29
East-Asia	Miao	10
East-Asia	Mongola	10
East-Asia	Naxi	10
East-Asia	Oroqen	10
East-Asia	She	10
East-Asia	Tu	10
East-Asia	Tujia	10
East-Asia	Yakut	25
East-Asia	Yi	10
Europe	Adygei	17
Europe	Basque	24
Europe	Bedouin & Druze[a]	95
Europe	French	29
Europe	Italian	13
Europe	Mozabite	30
Europe	Orcadian	16
Europe	Palestinian	51
Europe	Russian	25
Europe	Sardinian	28
Europe	Tundra Nentsi	14
Europe	Tuscan	8
Oceania	Melanesian	19
Oceania	Papuan	17

Table S2: Prior distributions and values of the parameters explored in the ABC analysis.

Parameter	Abbreviation	Distribution	Min.	Max.	Mean	S.D.
Initial Population Size	N_i	Uniform	2	5120	-	-
Carrying Capacity	K	Uniform	2	5120	-	-
Growth rate	r	Lognormal	0.01	2.5	0.5	0.6
Migration rate	m	Uniform	0	0.5	-	-
Time of the onset	T	Normal	2000	10400	6200	1280
Mutation rate	μ	Uniform	1.00E-05	1.00E-03	-	-

Chapter 2 – Natural selection in a post-glacial range expansion: the case of the colour cline in the European barn owl

Sylvain Antoniazza*, Ricardo Kanitz*, Samuel Neuenschwander, Reto Burri, Arnaud Gaigher, Alexandre Roulin, Jérôme Goudet

*co-first authors.

Manuscript under review in Molecular Ecology

Natural selection in a post-glacial range expansion: the case of the colour cline in the European barn owl

Sylvain Antoniazza,*[1] Ricardo Kanitz,*†[1] Samuel Neuenschwander,*‡ Reto Burri,§ Arnaud Gaigher,* Alexandre Roulin* & Jérôme Goudet*†

*Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

†Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

‡Vital-IT, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

§Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

[1] These authors contributed equally to this work

Keywords: approximate Bayesian computation, range expansion, natural selection, colour polymorphism, cline

Correspondence: Jérôme Goudet, Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland. Fax: +41 21 692 4265.

E-mail: jerome.goudet@unil.ch

Running title: Range expansion and a phenotypic cline

Abstract

Gradients of variation – or clines – have always intrigued biologists. Classically, they have been interpreted as the outcomes of antagonistic interactions between selection and gene flow. Alternatively, clines may also establish neutrally with isolation-by-distance or secondary contact between previously isolated populations. The relative importance of natural selection and these two neutral processes in the establishment of clinal variation can be tested by comparing genetic differentiation at neutral genetic markers and at the studied trait. A third neutral process, surfing of a newly arisen mutation during the colonisation of a new habitat, is more difficult to test. Here, we designed a spatially-explicit ABC simulation framework to evaluate whether the strong cline in the genetically-based reddish coloration observed in the European barn owl (*Tyto alba*) arose by allele surfing or whether selection has to be invoked to explain this colour cline, for which we have previously ruled out the actions of isolation-by-distance or secondary contact. Using ABC simulations and genetic data on 390 individuals from 20 locations genotyped at 22 microsatellites loci, we first determined how barn owls colonized Europe after the last glaciation. Using these results in new simulations on the evolution of the colour phenotype, and assuming various genetic architectures for the colour trait, we demonstrate that the observed colour cline cannot be due to the surfing of a neutral mutation. Taking advantage of spatially explicit ABC, which proved to be a powerful method to disentangle the respective roles of selection and drift in range expansions, we conclude that the formation of the colour cline observed in the barn owl must be due to natural selection.

Introduction

Determining the relative roles of natural selection and neutral processes as driving agents of evolutionary change has long been the focus of discussions in the field of evolutionary biology (Kimura 1983; Nei 2005; Wagner 2008). A process of particular interest in this context is one observed in many species presently occupying temperate areas: *range expansions*. Most (if not all) species currently inhabiting Europe and North America have undergone postglacial recolonisation events, increasing their ranges and population sizes (Hewitt 2000), and nowadays, some extant species and populations facing the on-going climatic changes and human alterations to the environment may also respond by increasing their range (Parmesan & Yohe 2003). Range expansions are a key factor for the discussion above because they often take place over an environmental gradient, which potentially provides natural selection with the opportunity to generate locally-adapted variants (Hewitt 1996). When these variants are distributed gradually across the environment, a cline is formed (Endler 1977). Clines along the path of range expansions, however, can also be formed without natural selection. The series of founder events, which are inherent to the colonization of new areas, may lead to the formation of allele frequency clines simply trough the neutral phenomenon of *allele surfing* (Edmonds *et al.* 2004; Klopfstein *et al.* 2006). In this process, neutral alleles may “surf” the wave of range expansion, increase their frequency along the way eventually forming a genetic cline. If the underlying genetics has any effect on phenotype, a purely neutral cline may become very similar to what one would expect to be a selection-derived cline (Currat *et al.* 2006).

Classically, clines have been studied in the context of hybrid zones, a secondary-contact zone between species or populations that evolved in allopatry, where selection against hybrids prevents gene flow and generates clines of phenotypes or alleles frequencies (Barton & Hewitt 1985). This is well described in the hybrid-zone literature, where the terms “cline” and “hybrid zone” are even sometimes confounded (Barton & Hewitt 1985). The processes behind the formation of such clines have been investigated in some details both theoretically and experimentally (Barton & Gale 1990; Barton & Hewitt 1985; Gay *et al.* 2008). Clines could also be the result of the mixing of populations adapted to different ecological conditions where the ecological transition occurs over short distances [e.g. latitudinal clines (James *et al.* 1997) or sharp environmental changes (Mullen & Hoekstra 2008)]. These ecological clines can be analysed in a similar way to the hybrid zones clines (e.g. Mullen & Hoekstra 2008). For these types of clines, the development of tools to infer selection has a long history and the method relies on the comparison between the clines’ width (w) and species’ dispersal distance (σ). In this case, selection is proportional to the square root of σ/w (Linnen & Hoekstra 2009; Slatkin 1973).

Clines can also appear through two neutral processes; isolation by distance and secondary contact without selective disadvantage of hybrids (Novembre & Di Rienzo 2009). When compared with natural selection, these neutral processes can essentially be ruled out by comparing the genetic/phenotypic variation putatively under selection to the neutral genetic variation. If the trait putatively under selection presents a stronger signal of population differentiation (higher Q_{ST} or P_{ST}) than neutral genetic markers (F_{ST}), there is probably selection involved in maintaining or leading to locally-adapted forms. Otherwise – if Q_{ST} is not significantly higher than F_{ST} – isolation by distance or secondary contact are enough to explain the observed patterns (Leinonen *et al.* 2008; Spitze 1993). Several studies have been performed to either compare differentiation at quantitative traits and neutral markers (Q_{ST}-F_{ST}) (Antoniazza *et al.* 2010; Demont *et al.* 2008; Gockel *et al.* 2001; Hangartner *et al.* 2012; Long & Singh 1995; Merilä 1997;

Palo *et al.* 2003; Savolainen *et al.* 2007; Storz 2002), or to compare genetic variation at different types of loci (F_{ST}-F_{ST}) (Ingvarsson *et al.* 2006; Kooyers & Olsen 2012; Saccheri *et al.* 2008).

In large-scale clines (occurring over wide geographical ranges, such as a continent), the role of selection has only been tackled through theoretical investigations focusing on either gene frequencies (Bazykin 1969; Endler 1977; Fisher 1950; Haldane 1948), or quantitative phenotypic traits (Barton 1999; Case & Taper 2000; Kirkpatrick & Barton 1997; Leimar *et al.* 2008; Slatkin 1978). No methods have yet been developed to infer selection in this case. In addition, the empirical studies describing large-scale clines have consistently neglected the evaluation of the surfing phenomenon as their possible cause. They have largely assumed natural selection to be driving force leading to the observed patterns [see Currat et al. (2006) and Vasemägi (2006) for critical reviews, and Kujala (2012) for an exception]. Nevertheless, the most probable source of allele surfing – i.e. range expansions – is common. Most species inhabiting temperate latitudes of both hemispheres spent the last glacial maximum (LGM) in refugia that were closer to the equator than their current distribution and then expanded their range after the last ice age (Hewitt 1999, 2000; Taberlet *et al.* 1998).

Evaluating how likely it is for a given cline to originate by allele surfing (relative to natural selection) is essential to understand its biological basis and can bring key insight on the more general discussion about the prevalence of selective processes in biological evolution. The establishment of clines by allele surfing in range expansions, however, is more difficult to rule out by means of Q_{ST}-F_{ST} comparisons than the other two neutral processes. Surfing mutations may also occur in the loci underlying the candidate trait (Klopfstein *et al.* 2006), leading to an inflated Q_{ST} when compared to other random loci's F_{ST}. In order to deal with this situation, one possible approach is to first infer/reconstruct the most likely demographic history for the taxon under investigation. This can be done using approximate Bayesian computation [ABC (Beaumont *et al.* 2002; Bertorelle *et al.* 2010; Csillery *et al.* 2010; Sunnaker *et al.* 2013)], where simulations with variable scenarios and demographic-parameter values are used to infer which parameters are closest to the observed genetic data, and whether the species has undergone a range expansion (Eriksson *et al.* 2012; Estoup *et al.* 2004; Estoup & Clegg 2003; Itan *et al.* 2009; Neuenschwander *et al.* 2008b; Warmuth *et al.* 2012). Second, using these estimated demographic parameters, a new round of neutral simulations is carried out, focusing this time on the phenotypic trait showing clinal variation. Taking advantage of many replicates, this procedure allows assessing the probability of the cline under investigation to have been generated by purely neutral processes (i.e. allele surfing in a range-expansion scenario). A similar approach was successfully used by Roux *et al.* (2012) in a different biological context (balancing selection in *Arabidopsis* spp.).

One striking example of clinal variation is provided by the south-west/north-east cline in colour of the European barn owl (*Tyto alba*) described by Roulin and colleagues (Roulin 2003; Roulin *et al.* 2009), and analysed along with neutral genetic markers in Antoniazza et al. (2010). Based on a comparison of the spatial variation of the colour with the neutral genetic diversity, the latter study revealed that the south-west/north-east colour cline is significantly steeper than population differentiation at neutral genetic markers measured in the same populations. Antoniazza et al. (2010) discussed the surfing hypothesis, but did not test it. A major characteristic of neutral genetic diversity in European barn owls is a decline from south-western (Iberian Peninsula) to north-eastern Europe (North-Eastern Germany to Serbia, Fig. S2). The likely origin of this genetic diversity decline is a series of bottleneck events during the post-glacial colonisation of northern Europe. Here, we investigate whether a post-glacial colonisation model is compatible with today's observed genetic diversity of the European barn owl, and investigate how likely it is for the colour cline to have arisen by allele surfing (as opposed to natural selection) during colonisation.

To reconstruct past and current demography of the European barn owl, a dataset of 390 individuals genotyped at 22 microsatellites coming from 20 sampling locations (Fig. S3) in Western Europe was analysed with spatially explicit simulations within an approximate Bayesian computation (ABC) framework. The observed patterns were compared to those generated with spatially explicit computer simulations using several plausible historical scenarios (Table 1) and 6-9 demographic parameters (Table 2). Based on observed genetic patterns, classical phylogeographic analyses and ecological knowledge of the species, a scenario consisting of a single colonisation from the Iberian Peninsula was hypothesized. As geographic variation in genetic diversity might arise by other processes than colonisation, we also tested scenarios with a south-west/north-east gradient of effective population size and extinction rate. Additionally, considering that many species were shown to have more than one glacial refugium (Taberlet *et al.* 1998), we looked at models with two glacial refugia in the Iberian Peninsula and in Greece. Finally, to control for the possibility that the patterns observed might not be derived from a colonisation process, several models without colonisation were tested as well.

Using the parameters obtained for the best-supported scenario for neutral genetic markers, we ran additional simulations to model the evolution of the colour trait. Different possible genetic architectures underlying the colour trait were investigated. For each one of these, we estimated the probability of generating a cline as steep as the one observed in the natural populations without selection.

Material and Methods

I. Sampling and molecular analyses

From 20 locations throughout Europe, a total of 390 barn owls were sampled by collaborators working in survey programs, recovery centres, and museums (Fig. 1). Genomic DNA was extracted from the basal 1 mm of breast feather quills, or from blood or muscles stored in 96% ethanol. Extractions were performed either on a BioSprint 96 extraction robot using the BioSprint 96 DNA blood kit or using the DNeasy blood and tissue kit, following the manufacturer's protocols (Qiagen, Hilden, Germany).

Population genetic statistics were estimated from genotypes obtained for 22 polymorphic microsatellite loci [(Ta-202, Ta-204, Ta-206, Ta-210, Ta-212, Ta-214, Ta-215, Ta-216, Ta-218, Ta-220, Ta-305, Ta-306, Ta-310, Ta-402, Ta-408 and Ta-413 from Burri *et al.* 2008) and (54f2, Calex-05, FEPO42, Oe053, GgaRBG18 and Tgu06 from Klein *et al.* 2009)]. Polymerase chain reactions (PCR) were performed in five multiplexes using the QIAGEN Multiplex PCR Kit (Qiagen, Hilden, Germany) and the following protocol: initial step of denaturation for 15 min at 95 °C, 34 cycles of 30 sec denaturation at 94 °C, annealing for 1.5 min at 57 °C, and elongation at 72 °C for 1 min. Final elongation for 30 min was conducted at 60 °C. The primer concentration and multiplexes composition can be found in Table S1. Fragment analyses were run on an ABI 3100 sequencer with a ROX 500 size standard and allele lengths were assigned using GENEMAPPER 4.0 (Applied Biosystems, Foster City, CA, USA). After verifying that no null-alleles were present (MICRO-CHECKER 2.2.3, Van Oosterhout *et al.* 2004) and that populations were not showing departure from Hardy-Weinberg equilibrium (Goudet 1995) the dataset was used to calculate observed summary statistics for the ABC estimation procedure. All summary statistics for both observed and simulated data were calculated using quantiNEMO (Neuenschwander *et al.* 2008a) and custom R scripts available on demand (R Development Core Team 2008).

II. Approximate Bayesian computation (ABC)

1. Population genetics patterns and choice of summary statistics

The rationale behind ABC is to compare simulated genetic data obtained under various scenarios and demographic/genetic parameters against observed genetic data through summary statistics (Beaumont 2010; Beaumont *et al.* 2002; Bertorelle *et al.* 2010; Csillery *et al.* 2010; Sunnaker *et al.* 2013). The choice of summary statistics on which the comparison is based is thus a key component of an ABC analysis. The summary statistics should describe the genetic data sufficiently, but should also be kept to a minimal number: Each additional summary statistic adds extra noise to the parameter estimation (Beaumont *et al.* 2002).

The present data exhibit strong geographic patterns of genetic diversity and population structure, which can be summarized by few summary statistics: (i) a significant signal of isolation-by-distance [IBD; pairwise F_{ST} as a function of pairwise geographic distances, Fig. S1 (Mantel test, $R^2 = 0.310$, $p < 0.001$)], and (ii) a significant reduction in genetic diversity from south-west to north-east [mean allelic richness per population as a function of geographic distance from the south-western most population, Fig. S2 ($R^2 = 0.779$, $p < 0.001$)]. Four statistics were implemented to summarize these patterns: (i) The IBD slope (5.68×10^{-4}); (ii) average mean pairwise F_{ST} between populations (1.68×10^{-2}); (iii) the slope of the regression of the mean allelic richness per population as a function of its distance to the south-western-most population (-2.18×10^{-2}); and (iv) the average mean allelic richness per population (5.32).

Table 1: Demographic models tested with ABC for the demographic history of the European barn owl. Two dimensions of the models are described (colonisation and heterogeneity) with the each model's number of variable parameters.

Colonisation model	**Heterogeneity model**	**Nb. of varying parameters (= with prior distributions)**
***One-refugium* (Iberian)**	*One-carrying-capacity* (base model)	6
	Carrying-capacity-cline (SW-NE)	7
	Extinction-rate-cline (SW-NE)	8
	Two-dispersal-rate (one during colonisation and one at carrying capacity)	7
***Two-refugium* (Iberian and Greek)**	*One-carrying-capacity*	7
	Carrying-capacity-cline (SW-NE)	8
	Extinction-rate-cline (SW-NE)	9
	Two-dispersal-rate (one during colonisation and one at carrying capacity)	8
No-colonisation	*One-carrying-capacity*	6
	Carrying-capacity-cline (SW-NE)	7
	Extinction-rate-cline (SW-NE)	8

2. Base model

One of the major drivers of barn owl populations' dynamics is winter harshness (Altwegg *et al.* 2006; Marti & Wagner 1985; Massemin & Handrich 1997). The sensitivity of this species to climate, notably to long periods of snow cover, is well known. There is no doubt that European barn owls endured the LGM in refugia in ice- and largely snow-free ranges south of their current European distribution. The strong cline in genetic diversity from south-west to north-east Europe points toward a single colonisation from the Iberian Peninsula (or north-Africa via Gibraltar, Fig. S2). Our basal simulation model is thus based on a colonisation of Europe from a single, Iberian glacial refugium.

Simulating colonisation processes requires spatially explicit modelling. A modified version of the quantiNEMO programme was used to simulate the colonization and the resulting neutral genetics (Neuenschwander *et al.* 2008a), using an integrated coalescent layer for increased efficiency. Our simulations consisted of two phases, similar to the approach implemented in SPLATCHE (Currat *et al.* 2004; Ray *et al.* 2010). In a first phase, spatially explicit demographic history was simulated forward in time (starting with the post-glacial colonization and ending today). In this phase the demographic history of populations is simulated based on the demographic parameters presented in Table 2. In the second phase genetic data were generated in a coalescent approach (backward in time, starting from today's sample and going back to the most recent common ancestor of all sampled lineages) using the demographic information obtained from the demographic simulations (Hudson 1990; Nordborg 2001). The genetic data (22 unlinked microsatellite markers) was simulated for the same number of individuals and populations as in the observed data. Mutations followed a stepwise mutation model (SMM).

Simulations were performed on a raster map of Europe consisting of 2671 square land demes, each 50 km × 50 km in size (Fig. S3). A deme may be inhabited by a single population. The deme size is appropriate for barn owl since their dispersal abilities are in this range (see below). Simulations started with a single population in the glacial refugium in the south of the Iberian Peninsula. At the start of a simulation, this refugium population was distributed in equal numbers among the nearest 100 demes (Fig. 1 and Fig. S3). In the following generations, the population range expanded successively across Europe based on demographic processes, such as local logistic population growth and dispersal to the four neighbouring demes (stepping stone dispersal model). This described base model requires five demographic parameters (time of the onset of colonisation, migration rate, deme carrying capacity, size of the refugium population, and intrinsic population growth rate) and a single genetic parameter (mutation rate of the microsatellites, Table 2).

3. Prior distributions

The prior distributions of the six base-model parameters for ABC analyses are based on extensive ecological data:

- *Start of the colonisation (time)*: As a result of high sensitivity of barn owls to winter harshness, the colonisation of the northern part of Europe necessarily occurred after the warming of the continent, i.e. after the LGM around 20 000 years ago (Clark *et al.* 2009). Since no information on the onset of colonisation is available, a broad uniform prior was chosen ranging from 2000 to 10 000 generations, which is about 7200-36 000 years BP assuming a constant generation time of 3.6 years for barn owls (Altwegg *et al.* 2006).

- *Dispersal rate*: Dispersal rate is generally high and with long ranges in the barn owl. In the Netherlands, more than 30% of the juveniles disperse more than 50 km from their place of birth to their place of reproduction (Bairlein 1985; Bunn *et al.* 1982). We account for these dispersal distances by defining a deme size of 50 km × 50 km (see above) and by defining a wide uniform dispersal prior allowing for high dispersal rates from 0 to 0.5.
- *Carrying capacity*: The barn owl census population size is well estimated in Europe and it counts about 140 000 breeding pairs (Hagemeijer & Blair 1997). We chose to cover a broad interval of 5-10 000 individuals per deme (so between 13 355 and 26 710 000 overall), but we put more weight on small values by using a lognormal distribution with a mean of 300 and a variance of 400.
- *Size of the refugium population*: As no information is available about this population size we used a wide uniform distribution between 100 and 100 000 individuals.
- *Population growth rate*: We chose a wide uniform prior between 0 and 2. Note, that the growth rate has only an effect during colonization when population size has not reached carrying capacity and is thus not a so important parameter of the model.
- *Mutation rate*: a lognormal distribution between 10^{-8} and 10^{-2} with a mean of 10^{-3} and a variance of 8×10^{-2} was used as prior to span the full range of plausible mutation rate values (Ellegren 2000).

Table 2: Demographic parameters for the different scenarios (models). Details on the *a priori* value distributions of the different models. Uniform distributions have equal probability of sampling any value between the defined boundaries; lognormal distributions have a higher probability of sampling values closer to its mean in a logarithmic scale, with predefined upper and lower limits (truncated). The brackets describing lognormal distributions give: (lower bound, upper bound, mean, variance).

Parameters	**For which model**	**Prior characteristics**
Start of the colonisation	All models	Uniform (2000-10 000 generations)
Population growth rate	All models	Uniform (0-2)
Mutation rate	All models	Lognormal ($1e^{-8}$-$1e^{-2}$, $1e^{-3}$, $8e^{-2}$)
Size of refugium population	All models	Uniform (100-100 000) but for 2-refugia models Uniform (200-100 000) for 1-refugium models
Dispersal rate	All models but *two-dispersal-rate*	Uniform (0-0.5)
Dispersal rate high density	*Two-dispersal-rate*	Uniform (0-0.5)
Dispersal rate low density	*Two-dispersal-rate*	Uniform (0-0.5)
Carrying capacity	All models but *carrying-capacity-cline*	Lognormal (5-10 000, 300, 400)
Carrying capacity of the SW deme	*Carrying-capacity-cline*	Lognormal (5-10 000, 300, 400)
Carrying capacity of the NE deme	*Carrying-capacity-cline*	Lognormal (5-10 000, 300, 400)
Extinction rate SW deme	*Extinction-rate-cline*	Uniform (0-0.5)
Extinction rate NE deme	*Extinction-rate-cline*	Uniform (0-0.5)
Divergence time	*Two-refugium*	Uniform (0-120 000)

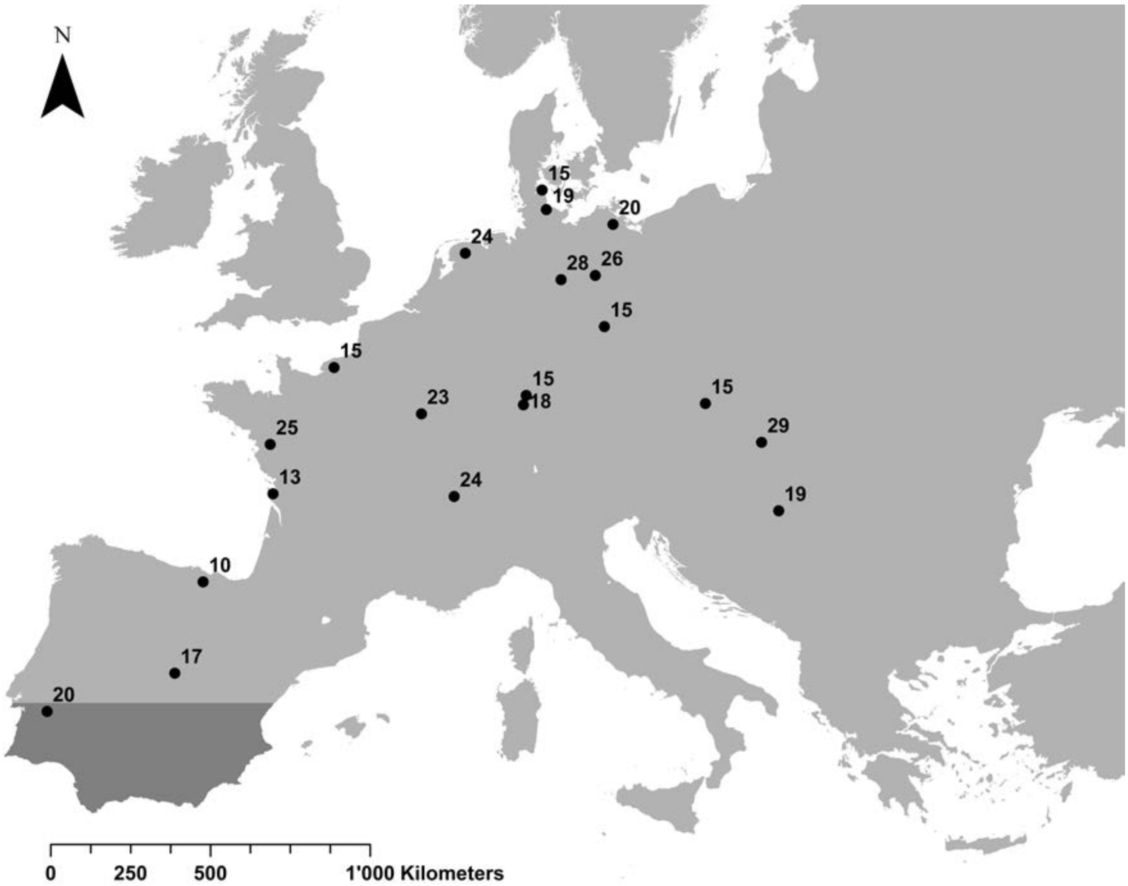

Figure 1: Map of the sampling locations and sampling sizes. Sampling sizes and sampling locations for the observed dataset are indicated. Similar sampling locations and sampling sizes are generated for the simulated dataset. The Iberian glacial refugium demes are indicated in dark grey. We use a Europe Albers Equal Area Conic projection to adequately represent surfaces.

4. Model comparison

ABC does not only allow estimating model parameters, but it is also effective in contrasting different models (eg. Sunnaker *et al.* 2013 and references therein). We took advantage of this feature to test for different scenarios that could explain the barn owl's post-glacial evolutionary history, and then applied more refined parameter estimation to the best-supported model.

Three colonisation models: Our observation of a strong decrease of allelic richness from the Iberian Peninsula towards northeastern populations (Fig. S2) suggests a single colonisation from the Iberian Peninsula. Our base model (*one-refugium* model) thus consists of a single colonisation of Western Europe from this Peninsula. However, many taxa in Europe are known to have survived the cold period also in eastern glacial refugia (Hewitt 1999; Taberlet *et al.* 1998). We tested this hypothesis by adding an additional eastern glacial refugium, of identical size situated in Greece, to previous model (*two-refugium* model). Finally, we tested the hypothesis if barn owls resisted the cold period and remained across Europe and thus had no colonisation phase after the LGM. We implemented this model by directly spreading the initial population size over the whole continent (*no-colonization* model).

Four heterogeneity models: The described base model has constant environmental characteristics (*one-carrying-capacity* model), i.e. deme characteristics did not change over space. However, several ecological characteristics of the barn owl, apart from the colonisation, might have induced spatial variation in genetic diversity. Half of the extant European barn owls are breeding in the Iberian Peninsula, and there is a strong decrease in population sizes from south-western to north-eastern Europe (Hagemeijer & Blair 1997). We thus tested whether a model with clinal variation in carrying capacity from south-west to north-east Europe fits the data better (*carrying-capacity-cline* model). A second key characteristic that might influence the spatial variation in allelic richness is the variation in the extinction rate. The European barn owl is very sensitive to cold, snow-rich winters, and the gradient of continentality from southwestern to northeastern Europe might play an important role in creating the observed pattern of genetic variation. We thus also ran a model that includes a south-west/north-east cline in extinction rates (*extinction-rate-cline* model). Finally, the last model investigated is based on the observation that migration rates may differ depending on the stage of colonisation looked at: Dispersal is often higher during the colonisation and then it lowers once carrying capacity is reached (Neuenschwander *et al.* 2008b; Saether *et al.* 1999). In this model (*two-dispersal-rate* model), we allowed for two migration rates, one at low density during colonization and one at high density when demes are completely populated.

The four heterogeneity models were combined with the three colonization models. The combination of the *no-colonization* and *two-dispersal-rate* models was eliminated since the migration rate during colonization would have no effect. Eleven different models were therefore compared (Table 1).

For the model comparison in ABC, we run 10^5 simulations for each of the eleven models based on parameters drawn from the corresponding prior distributions (Table 2). Each simulation was compared to the observation by their summary statistics, resulting in a Euclidean distance. Models were then compared based on their posterior probabilities following Leuenberger and Wegmann (2010) as implemented in ABCTOOLBOX (Wegmann *et al.* 2010).

5. Parameter estimates

The best demographic model was then selected for final parameter estimation. A total of 10^6 simulations were generated as before based on parameters drawn from the prior distributions. 1000 simulations closest to the observation were retained for parameter estimation using a locally-weighted linear-regression approach implemented in the package ABCTOOLBOX (Wegmann *et al.* 2010).

6. Quality assessment of estimates

To test the accuracy of our estimates, we use 1000 randomly chosen simulations (from the 10^6 simulations dataset) with known parameter values and their resulting genetic data as pseudo-observations. Using the same ABC framework as before, we estimated the parameter values for these pseudo-observations. The accuracy of the estimation was measured by comparing the estimated parameter value against the "true" parameter value using the following statistics: relative root mean square error (RMSE), mean relative bias, proportion of high posterior density 50% (HPD50%) encompassing the pseudo observed value, proportion of HPD95% encompassing the pseudo-observed value. We also computed the R^2 of the linear regression of the estimated parameter values as a function of the pseudo-observed parameter values (Neuenschwander *et al.* 2008b).

III. Simulations applied to the colour trait

Additional simulations were performed to assess the probability of neutral processes generating the colour cline observed in the barn owl across Europe. These simulations were run using the best demographic model and parameter values drawn from the posterior distributions (HPD95% intervals). In order to simulate colour as a quantitative trait, we ran the simulations forward in time in quantiNEMO (Neuenschwander et al. 2008).

The individual breast-colour variation in the barn owl ranges from purely white to rufous-brown (dark). Because the genetic basis for this trait is still poorly known (Roulin & Dijkstra 2003), we investigated five alternative genetic architectures. (i) The simplest architecture consists of a single bi-allelic locus; more complex ones involved (ii) 25 bi-allelic loci; and (iii) a single multi-allelic locus (with 50 alleles). For these three architectures, the determination of the colour phenotype was defined as purely additive (i.e. no dominance, nor epistasis). Additionally, we explored architectures with (iv) a single bi-allelic locus and (v) 25 bi-allelic loci, where the dark allele was completely recessive. Even though somewhat unrealistic, this dominance scheme was used in order to allow for a higher initial dark allele frequency in the refugium, while keeping the frequency of the dark phenotype at its observed value, which would facilitate the surfing phenomenon (Hofer *et al.* 2009). In other words, we chose this dominance scheme in order to be conservative, by favouring the neutral processes.

In the bi-allelic architectures (for either one locus or 25 loci), one allele was considered “white” (representing the whitest birds), the other “dark” (representing the darkest birds). In the multi-allelic architecture, alleles are distributed over a linear gradient ranging from “whitest” to “darkest” with 50 different levels. Also, as a control, we simulated 22 microsatellite loci to mimic the purely neutral markers used in the previous simulations.

As the initial frequency of a given allele (Hofer *et al.* 2009) or the geographic location where a new allele appears (Klopfstein *et al.* 2006; Travis *et al.* 2007) may play a major role in the probability of observing the surfing phenomenon, two models varying in these respects were designed: (i) evolution from standing variation and (ii) enforced allele surfing. This second scenario was implemented only for the architectures without dominance, in order to estimate the probability of surfing for a new mutation occurring at the front of the expansion. For all these models, range expansion started from an Iberian refugium colonising the rest of the continent, potentially generating clines in colour polymorphism through the process of allele surfing.

Initial allele frequencies depended on the model used. For models based on standing variation, the average initial frequencies were calculated based on the current phenotype frequencies observed in the refugium of the Iberian Peninsula where the white phenotype is currently present at a ~90% frequency. Accordingly, for the codominance models, the initial frequency of the "white" allele was 90%; for the complete dominance models, it was 68%. In the multi-allelic model, the frequency of each allele was given by an exponential distribution, in which the lighter-coloured half of the alleles had a frequency of 90%.

For the simulations with enforced allele surfing, the whole Iberian Peninsula started already occupied and the white allele was fixed in all patches. One patch, located in the north-eastern corner of the Iberian Peninsula, contained a single dark allele at each locus (also for the multi-locus architecture). For the multi-allelic trait, the new mutation was implemented by bringing the darkest allele into the population which, in this case, contained the same exponential distribution of the other alleles as used in the evolution from standing variation scenario. As a result, this dark allele was at the very front end of the expansion, giving it an enhanced chance to spread by hitchhiking on the colonisation wave and creating the observed cline.

Beyond dominance effects, the mapping of genotypes into phenotypes was also done considering two different values for heritability of the colour trait ($h^2 = 0.81$ or 1). These values were chosen because narrow-sense heritability for colour was estimated to be 0.81 in Switzerland (Roulin & Dijkstra 2003; Roulin *et al.* 1998), and complete heritability ($h^2 = 1$) makes the estimation of phenotypic differentiation (Q_{ST}) more conservative.

For all simulations, we calculated the linear regression between pairwise geographic distances and the neutral genetic (F_{ST}) or phenotypic differentiation (Q_{ST}) between the 20 sampled populations. To assess the steepness of the cline produced, we retained the slope of the linear regressions, and following (Antoniazza *et al.* 2010) used the difference in slope between Q_{ST} and F_{ST} as a statistic to summarize the discrepancy between phenotypic and neutral markers differentiation. Finally, we compared the values for the difference of slopes obtained in each one of the simulation models with the relative position of this statistic as calculated for the observation. The proportion of simulations in each model that returned values equal or higher than the observation provided us with an estimate of the probability of attaining the observed values with that given neutral model.

Results

Model comparison

The posterior probability of each of the eleven models tested is presented in Fig. 2. The four models with one glacial refugium are best supported and their total posterior probability is higher than 90%. Among the one-refugium models, the base scenario with a constant carrying capacity over the continent had the highest posterior probability (0.31), followed by the model with a south-west/north-east cline in carrying capacities (0.25). The former model has not only higher support, but is also more parsimonious than the latter and was therefore used for all further simulations. Interestingly, the estimation of the parameters for the second best model, with a cline in carrying capacity, (although clearly less supported) results in estimates for a very shallow or non-existents cline of carrying capacities supporting the best simple model (estimates not shown).

Demographic parameter estimates

The posterior distributions for the demographic and genetic parameters of the base model (*one refugium* with *one carrying capacity*) are shown in Fig. 3, and the corresponding point estimates are reported in Table 3. The carrying capacity shows a narrow posterior distribution with a mode at 203 individuals per deme and a HPD95%

varying between 76.5 and 555. The population growth rate, as well as the refugium population size, shows broad posterior distributions, and their point estimates of respectively 1.58 and 59 800 should be considered with caution given their low estimability (see below). Dispersal rate estimates show high values with a mode at 0.375 and an HPD95% of 0.188-0.5. The mutation rate showed a very narrow posterior distribution with a mode of 1.03×10^{-4} and a HPD95% between 2.85×10^{-5} and 3.8×10^{-4}. The estimation of the onset of colonisation indicates high values with a mode at 7350 generations, which corresponds to about 24 500 years BP according to the generation times estimated in a Swiss barn owl population (Altwegg *et al.* 2006) and its HPD95% varies between 3810 and 10 000 generations ago.

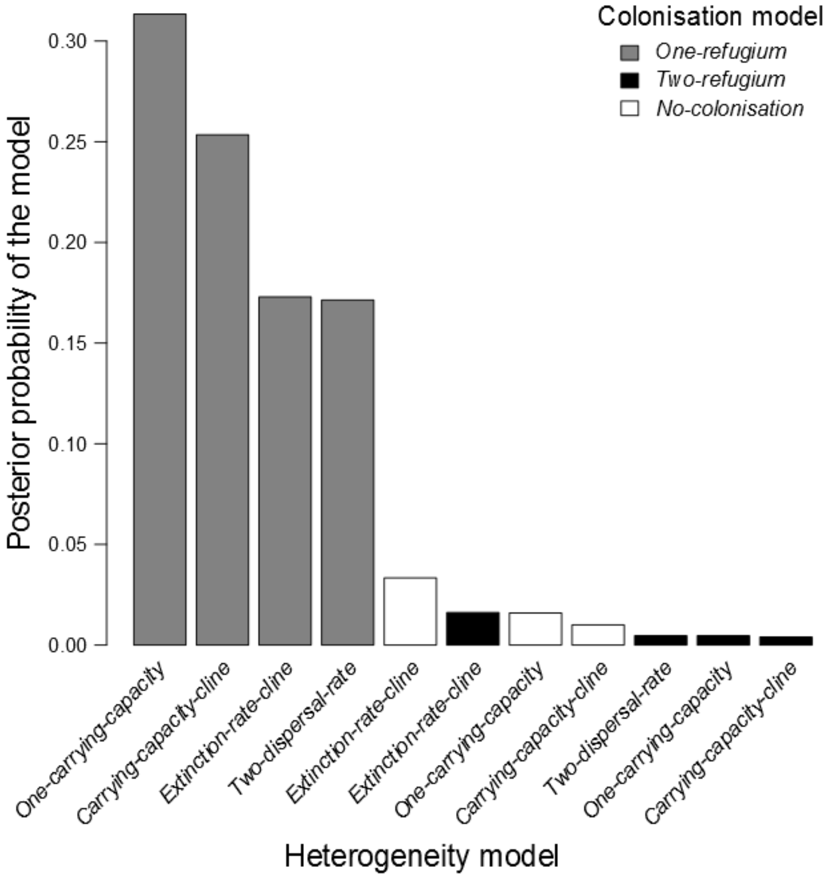

Figure 2: Posterior probabilities of the 11 models tested based on Leuenberger and Wegmann (2010). Based on four pattern statistics, 1000 simulations over 100 000 simulations per models were retained.

Table 3: Parameter estimates under the best-supported model (one-refugium single carrying capacity). Estimated modes are used as point estimates; HPD95% stands for the 95% highest posterior density intervals.

Demographic parameters	Estimated modes	HPD95%
Start of the colonisation (generations)	7350	3810 - 10 000
Population growth rate	1.58	0.338 - 2.00
Mutation rate	1.03×10^{-4}	2.85×10^{-5} - 3.80×10^{-4}
Size of refugium population	59 800	8840 - 98 800
Dispersal rate	0.375	0.188 - 0.500
Carrying capacity	203	76.5 - 555

Quality assessment

All statistics that assess the quality of the estimation of the parameters (RMSE, R^2, relative bias) are consistent with which parameters can be well estimated and which ones cannot (Table 4). The RMSE of the parameter estimation varies widely from 0.0516 to 7.55. While the estimation of mutation rate is highly accurate; dispersal rate, carrying capacity and start of the colonization are estimated less accurately. Population growth rate and refugium population size however show poor accuracy, which is not unexpected since these parameters have only an effect on the demographic history during a short period of the simulation (i.e. during the colonization process).

The validation analyses showed that our posteriors estimates are generally conservative (Table 4): More pseudo-observed simulations are generally found in the posterior distribution than expected (except for the start of colonisation and the refugium population size that show a slight deficiency). As those distributions were used as the background model for the colour simulations, we can be confident that they provided solid foundations.

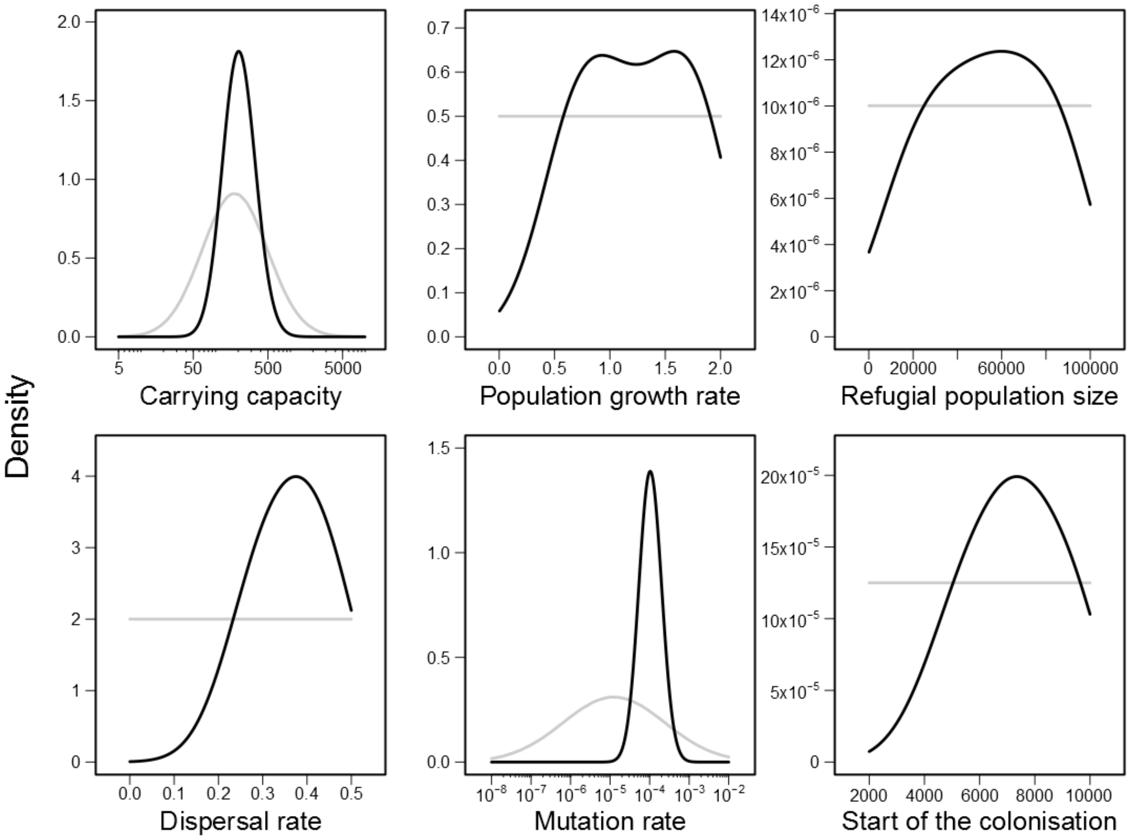

Figure 3: Posterior distributions of the estimated parameters. Grey lines show prior density and black lines the posterior distributions. Note that carrying capacity and mutation rate are in logarithmic scale.

Colour simulations reveal adaptive origin of colour cline

Lastly, the probability of generating the observed colour cline under a strictly neutral model was assessed with a second round of simulations. For each combination of genetic architecture and model of polymorphism distribution (dominance or not, enforced allele surfing or not), we generated 1000 replicates. The results for the comparison between these simulations and the observed values are presented in Fig. 4. For the models based on standing variation, we observe that no simulation reached the observed values (with either $h^2 = 1$ or 0.81), no matter the dominance. When mutations are enforced to take place at the very front of the expansion, between one and five simulations produced difference of slopes equal to or larger than what is observed in the owl populations (1 out of a 1000 for both one-locus traits, 3 and 5 out of 1000 for the multi-locus trait with the $h^2 = 0.81$ or $h^2 = 1$, respectively). In summary, without selection, very few simulations under an unlikely scenario managed to recreate the abrupt cline in colour visible in the observed data.

Table 4: Validation of the estimates for the *one-refugium*, *one-carrying-capacity* model based on 1000 pseudo-observations. See Material & Methods, *Quality assessment of estimates* for more details.

Demographic parameters	**Rel. bias**	**Rel. RMSE**	**Prop. HPD 50%**	**Prop. HPD 95%**	**R^2**
Start of colonisation	0.109	0.536	47	93.2	0.154
Population growth rate	0.744	5.9	52.6	94.9	0.0952
Mutation rate*	0.00336	0.0516	60.8	98.0	0.955
Refugium population size	1.24	7.55	48.9	94.7	0.0582
Dispersal rate	0.206	0.919	56.5	97.3	0.576
Carrying capacity*	0.00978	0.114	65.5	98.7	0.649

*As for the parameter estimate, these parameters are in log scale.

Discussion

Neutral demographic model

The spatially explicit approximate Bayesian computation analysis strongly supported the hypothesis that barn owls colonised Europe after the LGM from a single refugium situated on the Iberian Peninsula. It appears that the sequential bottlenecks during colonization alone explain the pronounced continuous decrease in diversity from the Iberian Peninsula to Eastern Europe. Alternative models including additional processes capable to explain the observed cline (cline in carrying capacity, cline in extinction rate, or hybrid zone due to two refugia) were worse than the simpler model. The parameter estimates of the demographic model for the European barn owl's post-glacial history inferred from neutral genetic data is consistent with what is known from other species (Hewitt 2000) and with the ecological literature on the species (see below).

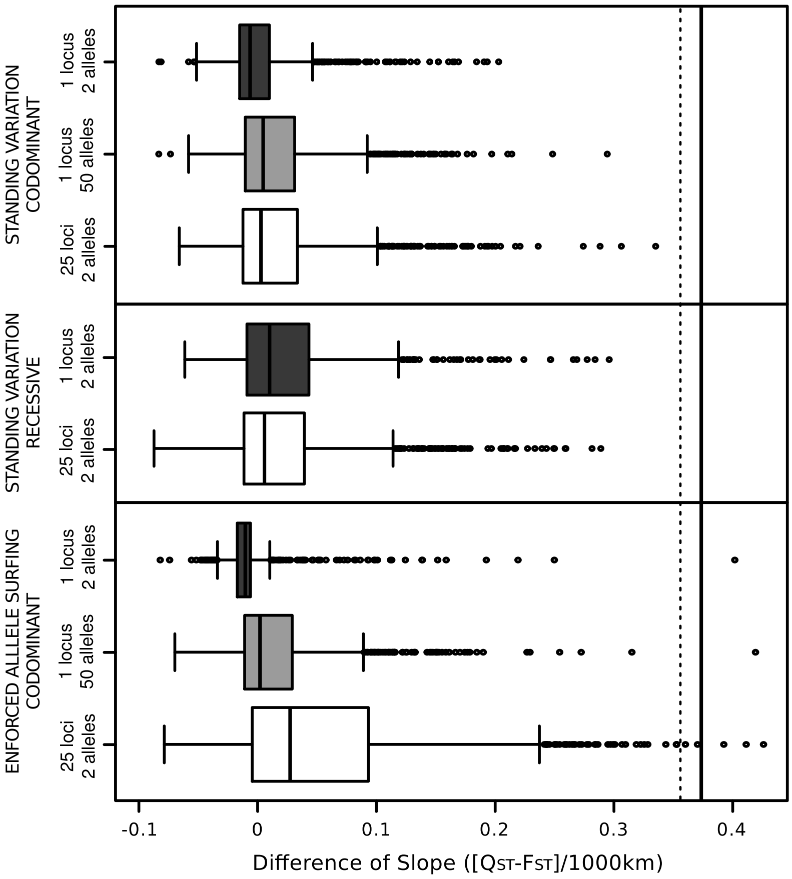

Figure 4: Probability of the neutral simulations to replicate the observed cline in colouration in the European barn owl. Comparison of distributions obtained from the calculation of the slope of IBD for the quantitative trait (colour, Q_{ST}) and the neutral loci (F_{ST}). Each model of different genetic architectures and starting polymorphisms is represented as a different distribution. The observed values for different levels of heritability are represented by the vertical lines: dashed line with $h^2 = 1$, plain line $h^2 = 0.81$.

The estimated start of the colonisation of 7350 generations ago with a 95% confidence interval of 3810 to 10 000 generations ago corresponds to 26 460 years BP and a 95% confidence interval from 36 000 to 13 700 year BP (assuming the estimated 3.6 years per generation, Altwegg et al. 2006) falls in line with an expansion following the LGM. This estimate is, although slightly higher, in good agreement with the expected time of colonisation of Europe after the LGM 20 000 years BP. The estimated time of colonization in years BP depends highly on the generation time, which is difficult to estimate and also assumes that the generation time remains constant over time.

The estimated carrying capacity of ~200 individuals per deme with a 95% confidence interval of 77 to 555 individuals extrapolated to the European scale results in a population size of 542 000 with a confidence interval of 204 300 to 1 482 400. Compared to the estimate of 140 000 breeding pairs (Hagemeijer & Blair 1997), i.e. 280 000 breeding individuals in Europe this seems to be overestimated. Even if these two numbers cannot be compared directly (the first one is an effective population size and the second one a census size for the breeding adults), they are in the same order of magnitude and our estimate is plausible. The overestimation is also based on the fact that the population size includes non-breeding individuals as well, such as juveniles [minimum age of the first breeding is one year (Cramp 1985)] and the simulated European map is slightly larger than the actual natural range of Barn owls.

The estimated migration rate of 0.375 between neighbouring demes of 50 km × 50 km is in accordance with what was estimated by Bunn et al. (1982) for the Netherlands (32.1 % of the young move more than 50 km in their first year). The estimated size of the refugium population of 59 800 individuals has to be taken with caution. As expected the accuracy tests show that this parameter is difficult to estimate since its traces in the genetic diversity in the present is secondary. In contrast, the estimate of the mutation rate is very accurate, with an estimate of 1.03×10^{-4} and a sharp 95% confidence interval from 2.85×10^{-5} to 3.80×10^{-4}. This estimate is in good agreement with the expectation (see discussion in Wegmann & Excoffier 2010).

To sum up, all estimates seem to be biologically meaningful and we are thus confident that this demographic and genetic model is a good approximation of the actual post-glacial history of the European barn owls and that it provides a sound demographic null model to investigate further questions regarding barn owl biology.

Colour simulations

Our finding that Europe was colonised from a single Iberian refugium has the implication that the colour cline (Fig. S4) might have been established by surfing during this colonisation. Simulations of a colour quantitative trait in our neutral demographic model were run in order to evaluate this possibility. Overall, the formation of the observed cline under neutrality is extremely unlikely: We only very rarely obtained simulations showing the same strong difference in geographic differentiation between phenotype and neutral markers. We never observed it with standing genetic variation in the refugium. Only with enforced allele surfing (i.e. explicitly seeding mutations in the front of expansion) we obtained between 1 and 5 simulations (out of a 1000) showing the same or larger differences. The highest number of such simulations was obtained using the trait architecture based on 25 bi-allelic loci (and hence, 25 mutations, 1 per locus, in the deme at the start of the expansion), but even with this unrealistically favourable architecture the probability was below 0.5%.

The colour simulations thus show that the evolution of the colour cline by surfing is extremely unlikely. The conclusion drawn by Antoniazza et al. (2010), that the European colour cline results from a local adaptation process, is thus confirmed by our simulation approach. With the exclusion of neutral scenarios, the evolution of the colour cline by natural selection generating local adaptation is indeed far more likely.

Evolution and maintenance of the European barn owl colour cline

The classical view on the evolution of barn owl colour variation in Europe is that the colour morphs evolved in allopatry in two refugia during the last glaciations and that the cline evolved by secondary contact after the ice age (Voous 1950). The model inferred above for the post-glacial history of the barn owl in Europe points toward a very different scenario. Our results suggest that the colour cline evolved during or after the colonisation out of a single refugium through a local adaptation process and also imply a very recent evolution of the colour cline (post-glacial, hence younger than 20 000 years BP). A rapid colonisation of Europe after the last ice age is supported by the observation of barn owl remains that dated at least from 10 000 years BP found in the UK (Del Hoyo *et al.* 2000; Yalden & Albarella 2009) and the estimated onset of colonisation of 7349 generation points toward a colonisation date close to the end of the last glacial maximum. Evaluating the strength of selection will be a next step to further the understanding of this system, but the lack of information on the selective agent behind the colour variation puts a serious challenge to this extension (Antoniazza *et al.* 2010).

Continental clines and evolution during range expansions

We feel that the case of the barn owl, where evolution of a locally adapted trait happened during or after the recolonisation of the continent after the ice age, might be far more common than currently recognized in other taxa. The climatic oscillations of the quaternary that shape the dynamics of the ranges of many species of temperate latitudes on both hemispheres, generated retreat/recolonisation cycles that occurred along major climatic axis (mainly north-south). Also, there is a growing body of evidence that local adaptation along such climatic axes is ubiquitous [e.g. size clines first describe by Bergmann (1847)].

Continental clines in temperate latitude thus offer a scope to study both local adaptation at large scale, but also the dynamics of this adaption in time and its interaction with colonisation processes. The interaction between natural selection and colonisation processes is a key question in evolutionary biology, but is still in its infancy (Excoffier *et al.* 2009). The study of large scale continental gradient might represent a fruitful area to study these questions in more details (see Kujala & Savolainen 2012 for a first approach with a non-spatial demographic model).

The European barn owl is a good illustration and provides a superb case study to investigate these questions. In this paper, we were able to show that the colour cline observed in this species was not established by neutral demographic processes during the colonisation of the European continent. This shows that selection processes must have been involved in the establishment of the European colour cline, even if the mechanism by which these colour clines established remain to be elucidated. The demographic model developed in this study will provide a sound neutral model for background process in the genome and be a solid starting point to tackle further evolutionary questions.

Acknowledgements

We would like to thank R. Alonso, A. Beaufils, C. Blaize, G. Burneleau, J. De Jong, K. Dichmann, J. Frank, J. Gonin, K. H. Graef, P. Iseli, H. Keil, A. Klein, F. Krause, O. Lambert, A. Marques, H. D. Martens, K. Poprach, R. Riep, I. Roque, M. Ruzic, H. Seeler, J. Sihelnik, J. Soufflot, F. Steiner, R. Tomé, I. Tórizs, F. Ziemann and I. Zuberogoïta for contributing with samples from many location across Europe. The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. Funding was provided by the Swiss National Science Foundation (SNSF) grants No. 31003A_120517 to AR and 31003A_138180 to JG.

Data Accessibility

Colour-phenotype data and microsatellite genotypes are available on Dryad (accession No. XXXXXX).

Author Contributions

SA and RK were responsible for the execution of this study: SA more focused on the demographic analyses; RK, on the quantitative-trait evolution analyses. Samples were collected by RB, SA and AR, and data were produced by SA, RB, AG and RK. SN contributed with analytical tools and participated in the analyses. JG and AR designed the research. RK, SA and JG wrote the main body of the paper and all authors contributed comments and editing on the final version of the manuscript.

Supporting Information

Supporting information is available further bellow and contains four figures (Fig. S1-4) and one table (Table S1).

References

Altwegg R, Roulin A, Kestenholz M, Jenni L (2006) Demographic effects of extreme winter weather in the barn owl. *Oecologia* **149**, 44-51.

Antoniazza S, Burri R, Fumagalli L, Goudet J, Roulin A (2010) Local adaptation maintains clinal variation in melanin-based coloration of european barn owls (*Tyto alba*). *Evolution* **64**, 1944-1954.

Bairlein F (1985) Dismigration und sterblichkeit in süddeutschland beringter schleiereulen (Tyto alba). *Vogelwarte* **33**, 81-108.

Barton NH (1999) Clines in polygenic traits. *Genetical Research* **74**, 223-236.

Barton NH, Gale KS (1990) Genetic analysis of hybrid zones. In: *Hybrid Zones and the Evolutionary Process* (ed. Harrison RG), pp. 13-45. Oxford University Press, Oxford.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics* **16**, 113-148.

Bazykin AD (1969) Hypothetical mechanism of speciation. *Evolution* **23**, 685-687.

Beaumont MA (2010) Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics* **41**, 379-406.

Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035.

Bergmann C (1847) Über die Verhältnisse der Wärmeökonomie der Thiere zu ihrer Grösse. *Göttinger Studien* **3**, 595-708.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology* **19**, 2609-2625.

Bunn DS, Warburton AD, Wilson RD (1982) *The barn owl* T. & A. D. Poyser, London.

Burri R, Antoniazza S, Siverio F, *et al.* (2008) Isolation and characterization of 21 microsatellite markers in the barn owl (*Tyto alba*). *Molecular Ecology Resources* **8**, 977-979.

Case TJ, Taper ML (2000) Interspecific competition, environmental gradients, gene flow, and the coevolution of species' borders. *American Naturalist* **155**, 583-605.

Clark PU, Dyke AS, Shakun JD, *et al.* (2009) The Last Glacial Maximum. *Science* **325**, 710-714.

Cramp S (1985) *The birds of the Western Palearctic: Terns to Woodpeckers.* Oxford University Press, Oxford.

Csillery K, Blum MGB, Gaggiotti OE, Francois O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution* **25**, 410-418.

Currat M, Excoffier L, Maddison W, *et al.* (2006) Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens" and "microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". *Science* **313**, 172.

Currat M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes* **4**, 139-142.

Del Hoyo J, Elliott A, Sargatal J (2000) *Handbook of the Birds of the World. Barn Owls to Hummingbirds* Lynx Editions, Barcelona.

Demont M, Blanckenhorn WU, Hosken DJ, Garner TWJ (2008) Molecular and quantitative genetic differentiation across Europe in yellow dung flies. *J Evol Biol* **21**, 1492-1503.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 975-979.

Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* **16**, 551-558.

Endler J (1977) *Geographic Variation, Speciation, and Clines* Princeton University Press, Princeton.

Eriksson A, Betti L, Friend AD, *et al.* (2012) Late Pleistocene climate change and the global expansion of anatomically modern humans. *Proceedings of the National Academy of Sciences* **109**, 16089-16094.

Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet JM (2004) Genetic analysis of complex demographic scenarios: Spatially expanding populations of the cane toad, Bufo marinus. *Evolution* **58**, 2021-2036.

Estoup A, Clegg SM (2003) Bayesian inferences on the recent island colonization history by the bird Zosterops lateralis lateralis. *Molecular Ecology* **12**, 657-674.

Excoffier L, Foll M, Petit RJ (2009) Genetic Consequences of Range Expansions. *Annual Review of Ecology Evolution and Systematics* **40**, 481-501.

Fisher RA (1950) Gene frequencies in a cline determined by selection and diffusion. *Biometrics* **6**, 353-361.

Gay L, Crochet PA, Bell DA, Lenormand T (2008) Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution* **62**, 2789-2806.

Gockel J, Kennington WJ, Hoffmann A, Goldstein DB, Partridge L (2001) Nonclinality of molecular variation implicates selection in maintaining a morphological cline of *Drosophila melanogaster*. *Genetics* **158**, 319-323.

Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F-statistics. *Journal of Heredity* **86**, 485-486.

Hagemeijer W, Blair MJ (1997) *The EBCC atlas of European Breeding Birds: their distribution and abundance.* T & AD Poyser, London.

Haldane JBS (1948) The theory of a cline. *Journal of Genetics* **48**, 277-284.

Hangartner S, Laurila A, Raesaenen K (2012) Adaptive divergence in moor frog (*Rana arvalis*) populations along an acidification gradient: inferences from Q_{ST}-F_{ST} correlations. *Evolution* **66**, 867-881.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. 247-276.

Hewitt GM (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society* **68**, 87-112.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-913.

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. *Annals of Human Genetics* **73**, 95-108.

Hudson RR (1990) Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology* (eds. Futuyma DJ, Antonovics J), pp. 1-44. Oxford University Press, Oxford.

Ingvarsson PK, Garcia MV, Hall D, Luquez V, Jansson S (2006) Clinal variation in phyB2, a candidate gene for day-length-induced growth cessation and bud set, across a latitudinal gradient in European aspen (*Populus tremula*). *Genetics* **172**, 1845-1853.

Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The Origins of Lactase Persistence in Europe. *Plos Computational Biology* **5**, e1000491.

James AC, Azevedo RBR, Partridge L (1997) Genetic and environmental responses to temperature of Drosophila melanogaster from a latitudinal cline. 881-890.

Kimura M (1983) *The Neutral Theory of Molecular Evolution* Cambridge University Press, Cambridge.

Kirkpatrick M, Barton NH (1997) Evolution of a species' range. *American Naturalist* **150**, 1-23.

Klein A, Horsburgh GJ, Kuepper C, *et al.* (2009) Microsatellite markers characterized in the barn owl (Tyto alba) and of high utility in other owls (Strigiformes: AVES). *Molecular Ecology Resources* **9**, 1513-1519.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol* **23**, 482-490.

Kooyers NJ, Olsen KM (2012) Rapid evolution of an adaptive cyanogenesis cline in introduced North American white clover (Trifolium repens L.). *Molecular Ecology* **21**, 2455-2468.

Kujala ST, Savolainen O (2012) Sequence variation patterns along a latitudinal cline in Scots pine (Pinus sylvestris): signs of clinal adaptation? *Tree Genetics & Genomes* **8**, 1451-1467.

Leimar O, Doebeli M, Dieckmann U (2008) Evolution of phenotypic clusters through competition and local adaptation along an environmental gradient. *Evolution* **62**, 807-822.

Leinonen T, O'Hara RB, Cano JM, Merilä J (2008) Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *J Evol Biol* **21**, 1-17.

Leuenberger C, Wegmann D (2010) Bayesian Computation and Model Selection Without Likelihoods. *Genetics* **184**, 243-252.

Linnen CR, Hoekstra HE (2009) Measuring natural selection on genotypes and phenotypes in the wild. *Cold Spring Harb Symp Quant Biol* **74**, 155-168.

Long AD, Singh RS (1995) Molecules versus morphology - the detection of selection acting on morphological characters along a cline in *Drosophila melanogaster*. *Heredity* **74**, 569-581.

Marti CD, Wagner PW (1985) Winter mortality in common barn-owls and its effect on population-density and reproduction. *Condor* **87**, 111-115.

Massemin S, Handrich Y (1997) Higher winter mortality of the Barn Owl compared to the Long-eared Owl and the Tawny Owl: Influence of lipid reserves and insulation? *Condor* **99**, 969-971.

Merilä J (1997) Quantitative trait and allozyme divergence in the Greenfinch (*Carduelis chloris*, Aves: Fringillidae). *Biological Journal of the Linnean Society* **61**, 243-266.

Mullen LM, Hoekstra HE (2008) Natural selection along an environmental gradient: A classic cline in mouse pigmentation. *Evolution* **62**, 1555-1569.

Nei M (2005) Selectionism and Neutralism in Molecular Evolution. *Mol Biol Evol* **22**, 2318-2342.

Neuenschwander S, Hospital F, Guillaume F, Goudet J (2008a) quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. *Bioinformatics* **24**, 1552-1553.

Neuenschwander S, Largiader CR, Ray N, *et al.* (2008b) Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. *Molecular Ecology* **17**, 757-772.

Nordborg M (2001) Coalescent theory. In: *Handbook of Statistical Genetics* (ed. Balding D). John Wiley & Sons, Chichester.

Novembre J, Di Rienzo A (2009) Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics* **10**, 745-755.

Palo JU, O'Hara RB, Laugen AT, *et al.* (2003) Latitudinal divergence of common frog (*Rana temporaria*) life history traits by natural selection: evidence from a comparison of molecular and quantitative genetic data. *Molecular Ecology* **12**, 1963-1978.

Parmesan C, Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. 37-42.

R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ray N, Currat M, Foll M, Excoffier L (2010) SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics* **26**, 2993-2994.

Roulin A (2003) Geographic variation in sexual dimorphism in the barn owl *Tyto alba*: a role for direct selection or genetic correlation? *Journal of Avian Biology* **34**, 251-258.

Roulin A, Dijkstra C (2003) Genetic and environmental components of variation in eumelanin and phaeomelanin sex-traits in the barn owl. *Heredity* **90**, 359-364.

Roulin A, Richner H, Ducrest AL (1998) Genetic, environmental, and condition-dependent effects on female and male ornamentation in the barn owl *Tyto alba*. *Evolution* **52**, 1451-1460.

Roulin A, Wink M, Salamin N (2009) Selection on a eumelanic ornament is stronger in the tropics than in temperate zones in the worldwide-distributed barn owl. *J Evol Biol* **22**, 345-354.

Roux C, Pauwels M, Ruggiero M-V, *et al.* (2012) Recent and ancient signature of balancing selection around the S-locus in Arabidopsis halleri and A. lyrata. *Mol Biol Evol*.

Saccheri IJ, Rousset F, Watts PC, Brakefield PM, Cook LM (2008) Selection and gene flow on a diminishing cline of melanic peppered moths. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 16212-16217.

Saether BE, Engen S, Lande R (1999) Finite metapopulation models with density-dependent migration and stochastic local dynamics. *Proceedings of the Royal Society of London Series B-Biological Sciences* **266**, 113-118.

Savolainen O, Pyhajarvi T, Knurr T (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution and Systematics* **38**, 595-619.

Slatkin M (1973) Gene Flow and Selection in a Cline. *Genetics* **75**, 733-756.

Slatkin M (1978) Spatial Patterns in Distributions of Polygenic Characters. *Journal of Theoretical Biology* **70**, 213-228.

Spitze K (1993) Population Structure in *Daphnia obtusa*: Quantitative Genetic and Allozymic Variation. *Genetics* **135**, 367-374.

Storz JF (2002) Contrasting patterns of divergence in quantitative traits and neutral DNA markers: analysis of clinal variation. *Molecular Ecology* **11**, 2537-2551.

Sunnaker M, Busetto AG, Numminen E, *et al.* (2013) Approximate Bayesian Computation. *Plos Computational Biology* **9**.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453-464.

Travis JMJ, Muenkemueller T, Burton OJ, *et al.* (2007) Deleterious mutations can surf to high densities on the wave front of an expanding population. *Mol Biol Evol* **24**, 2334-2343.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4**, 535-538.

Vasemägi A (2006) The adaptive hypothesis of clinal variation revisited: Single-locus clines as a result of spatially restricted gene flow. *Genetics* **173**, 2411-2414.

Voous KH (1950) On the distributional and genetical origin of the intermediate population of the barn owl (*Tyto alba*) in Europe. In: *Syllegomena biologica* (eds. Jordans A, Peus F), pp. 420-443. Akad. Verlagsgesellshaft Ziemsen Verlag, Leipzig Wittenberg.

Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* **9**, 965-974.

Warmuth V, Eriksson A, Bower MA, *et al.* (2012) Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 8202-8206.

Wegmann D, Excoffier L (2010) Bayesian Inference of the Demographic History of Chimpanzees. *Mol Biol Evol* **27**, 1425-1435.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC BIOINFORMATICS* **11**, 116.

Yalden DW, Albarella U (2009) *The History of British Birds* Oxford University Press, Oxford.

Supplementary Figures

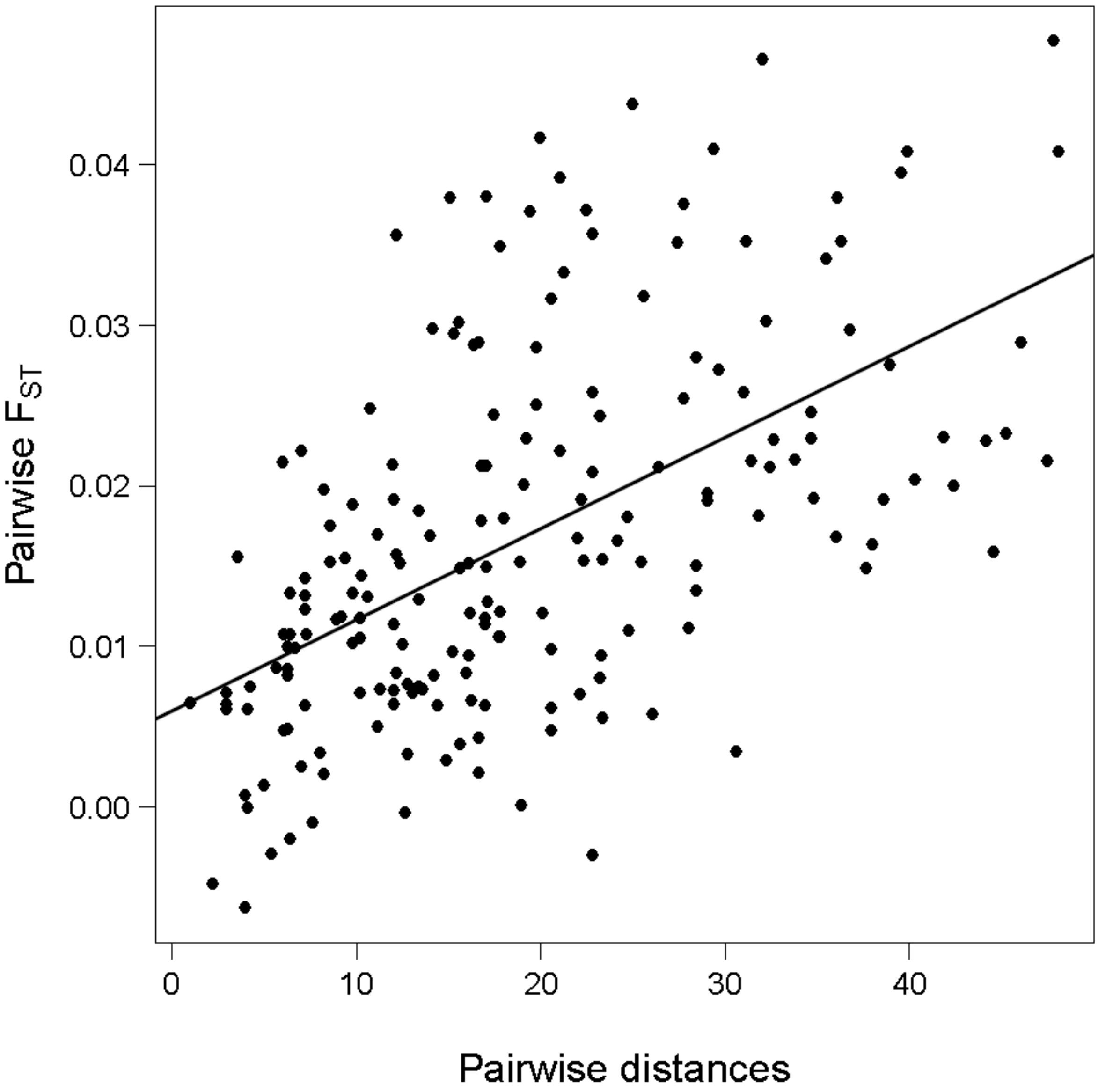

Figure S1: Observed isolation by distance (pairwise F_{ST} as a function of pairwise distances). Note that pairwise distances are in deme units (50 km). $R^2 = 0.310$.

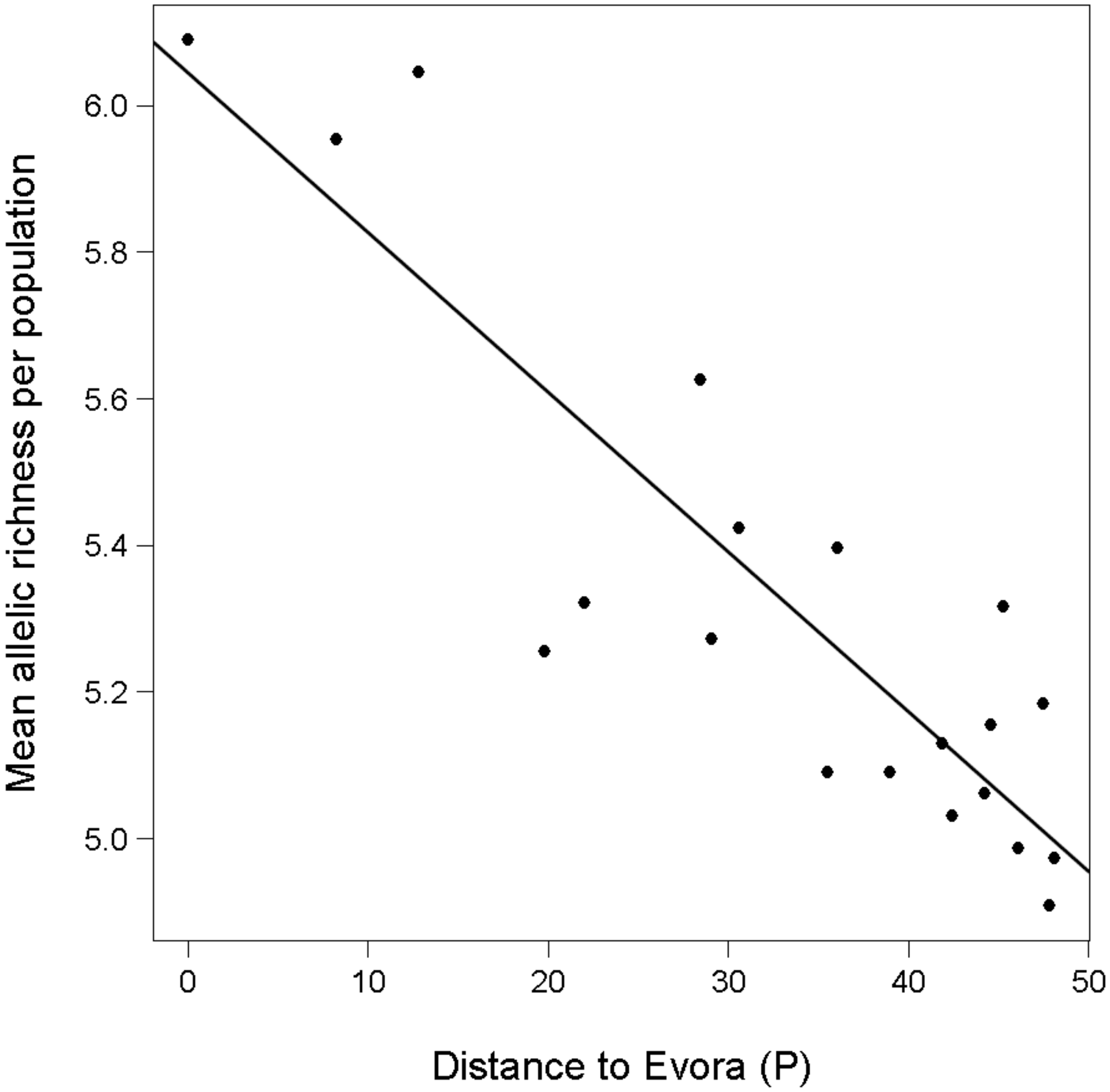

Figure S2: Observed cline in mean allelic richness (mean allelic richness over 22 loci per population as a function of distance to the south-westernmost population). Note that distances are in deme units (50 km). $R^2 = 0.779$.

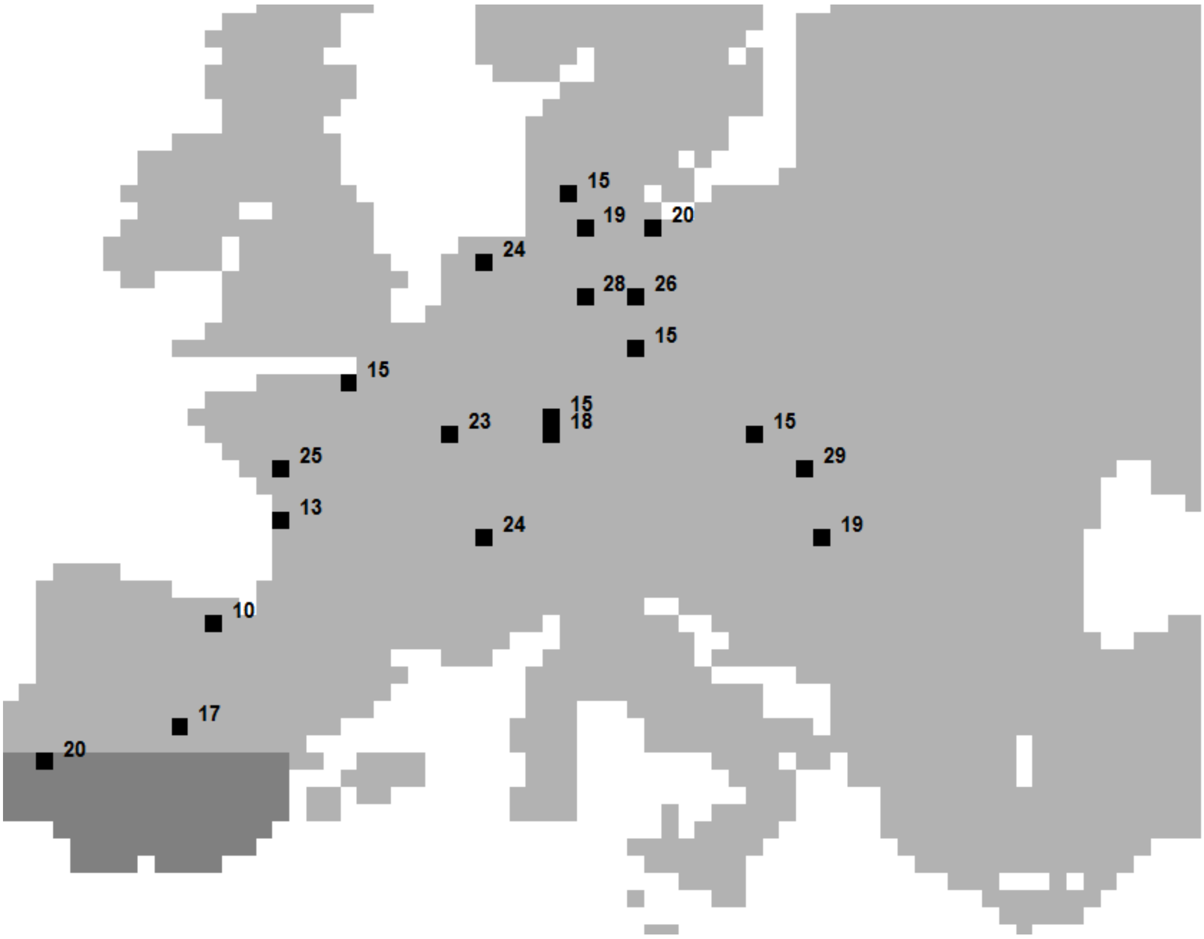

Figure S3: Simulated map of the sampling locations and sampling sizes. Sampling sizes and sampling locations as for the observed dataset are indicated. Similar sampling locations and sampling sizes are generated for the simulated dataset. Colonisable demes are indicated in grey, sea demes (not colonisable) are indicated in white. The Iberian glacial refugium demes are indicated in dark grey. We use a Europe Albers Equal Area Conic projection to adequately represent surfaces.

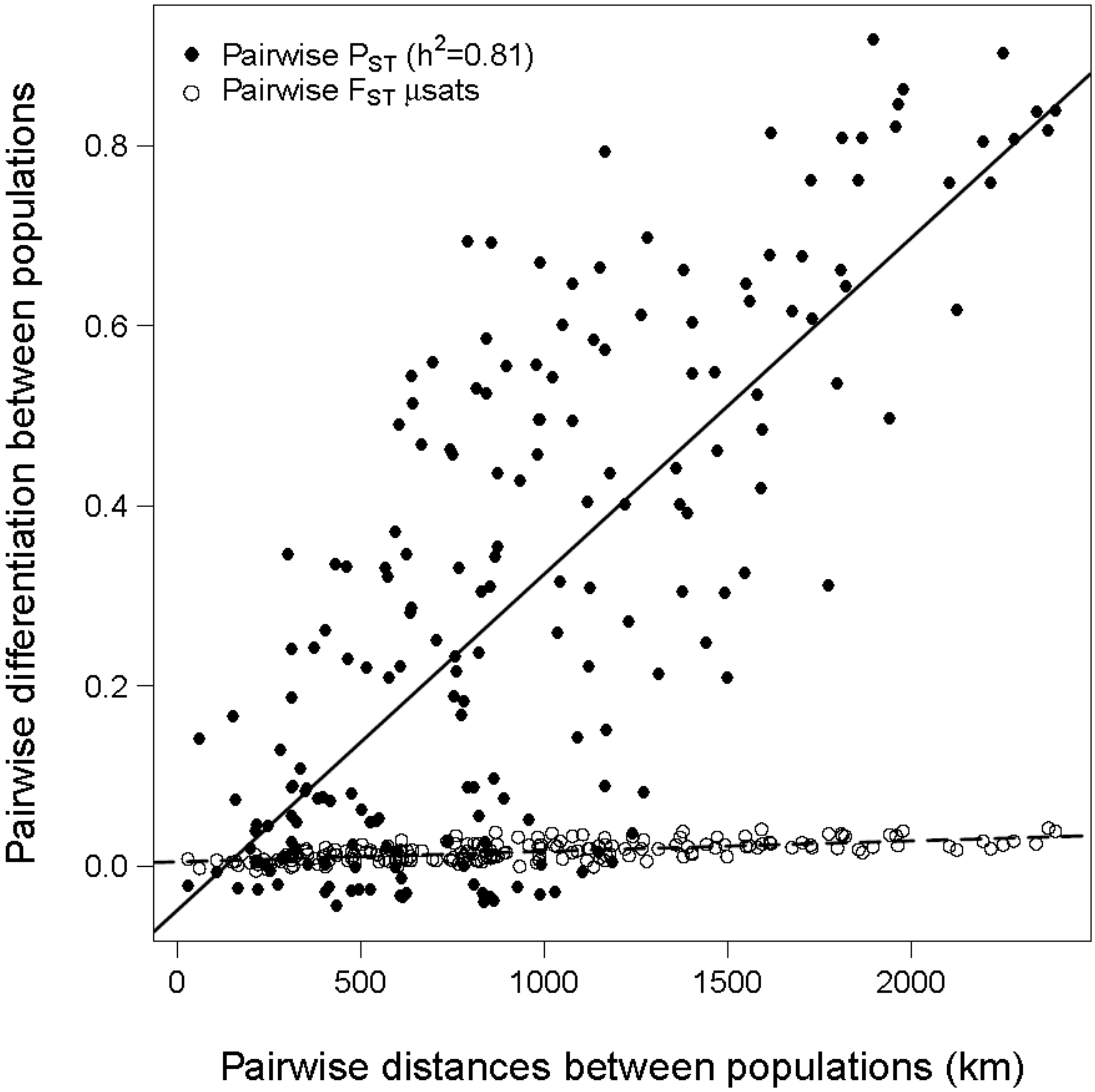

Figure S4: Observed isolation by distance for the microsatellites (pairwise F_{ST} as a function of pairwise distances, same data as in Fig. S1) and for the colour data (pairwise P_{ST} as a function of pairwise distances). Note that distances are in kilometres.

Supplementary Table

Table S1: Multiplex composition and primer concentration for microsatellite genotyping.

Multiplex	Locus	Dye	Final Conc. [μM]
Multiplex 1	Ta-206	FAM	0.45
	Ta-210	HEX	0.105
	Ta-216	FAM	0.135
	Ta-306	NED	0.165
Multiplex 2	Ta-218	HEX	0.178
	Ta-220	FAM	0.11
Multiplex 3	Ta-204	HEX	0.25
	Ta-214	FAM	0.5
	Ta-305	FAM	0.5
	Ta-310	NED	0.25
	Ta-413	NED	0.25
Multiplex 4	Ta-202	FAM	0.25
	Ta-212	DYO630	1
	Ta-215	FAM	1
	Ta-402	NED	0.25
	Ta-408	HEX	0.5
Multiplex 5	FEPO42	FAM	0.24
	54f2	NED	0.24
	Tgu06	HEX	0.48
	Calex-05	DYO630	0.48
	RBG18	FAM	0.72
	Oe053	HEX	0.96

Primer concentration is indicated for both forward and reverse primer together.

Chapter 3 – Natural selection in range expansions: insights from a spatially explicit ABC approach

Ricardo Kanitz, Samuel Neuenschwander, Jérôme Goudet

Manuscript to be submitted to Genetics

Natural selection during range expansions: insights from a spatially explicit ABC approach

Ricardo Kanitz[1,2,*], Samuel Neuenschwander[1,3], Jérôme Goudet[1,2,*]

[1]Department of Ecology & Evolution, University of Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, University of Lausanne, Switzerland

[3]Vital-IT, Swiss Institute of Bioinformatics, University of Lausanne, Switzerland

*Correspondence to ricardo.kanitz@unil.ch, jerome.goudet@unil.ch

Keywords: range expansion, natural selection, allele surfing, simulations, approximate Bayesian computation.

Running title: Natural selection in range expansions

Abstract

For at least 40 years now, evolutionary biologists have been discussing the relative roles of natural selection and genetic drift in shaping the genetic composition of populations. Range expansions are of particular interest in this discussion: They normally occur over environmental gradients allowing local adaptation to take place, but the demographic properties of these expansions also potentiate genetic-drift effects, which may in turn generate extreme changes in allele frequencies as populations expand in territory and numbers (i.e. allele surfing). Here, we address the detection and measurement of selection in such scenario using simulations. We mimic a range expansion over a variable selective gradient where individuals have in their genomes both loci that are neutral and loci determining a quantitative trait subject to selection. The responsiveness of summary statistics to the selective pressure is then assessed, and estimates of the selective pressure are made – based on these statistics – with approximate Bayesian computation (ABC). We observe that statistics related to isolation-by-distance patterns present a strong response to selection. This response can be used in ABC to estimate the strength of selection acting on the simulated populations with very reliable measures of estimability, regardless of the genetic architecture underlying the selected phenotypic trait. Furthermore, these estimates are robust to noise produced by genetic and demographic parameters such as heritability, mutation, migration and population-growth rates. This approach of taking into account the spatial dimension of differentiation in quantitative traits offers a promising avenue for investigating the role of natural selection in range-expansion scenarios, with possible implementations in the study of natural cases, as well.

Introduction

The opposition between selectionism and neutralism is one of the most important debates in evolutionary biology (EWENS 1977; KIMURA 1984; HEY 1999; NEI 2005). Ultimately, the question relies on which kind of process (neutral or selective) leads to the patterns observed in nature. Even though reconciliatory ideas have been proposed (WAGNER 2008), the dilemma regarding selection vs. neutrality still endures in different contexts of evolutionary biology (NEI *et al.* 2010). One evolutionary context that has drawn increasing attention from evolutionary biologists is the context of 'range expansions'. Range expansions are a ubiquitous phenomenon in nature involved in processes such as biological invasions (PARMESAN and YOHE 2003; WALTHER *et al.* 2009), adaptive radiations (RUNDELL and PRICE 2009), speciations (THORPE 1984; HEWITT 1996), pest and disease outbreaks (JEPSEN *et al.* 2008; ROTH *et al.* 2010), and post-glacial recolonizations (HEWITT 1996). Contractions and recolonizations following glacial oscillations are immensely common in nature, not only in temperate areas, but in tropical and subtropical regions, as well (COLINVAUX *et al.* 2000; HEWITT 2000). Therefore, it is safe to say that range expansions are likely involved in the evolutionary history of most of the organisms on the planet.

In the *selection vs. neutrality* discussion, range expansions are particularly important because populations increasing their range tend to do so over environmental gradients, leaving room for selection to act, possibly leading to local adaptation (HEWITT 1996). When different forms are established across this gradient, a *cline* is produced (ENDLER 1977). Clines have been thoroughly studied in the context of hybrid zones, where two allopatric populations get into secondary contact forming a tension zone in which the hybrids are selected against, so that the width of the cline is inversely proportional to the strength of selection (BARTON and HEWITT 1985). The same rationale was later applied to clines appearing in ecological transition zones (i.e. ecotones): MULLEN and HOEKSTRA (2008), in what has become a classical example, demonstrated that strong selection maintains two color-morphs of deer mice separated in two different habitats. These studies, however, have concentrated on small geographical scale clines. When it comes to large-scale clines (such as those appearing across continents) the literature is scarcer with some theoretical studies focused on gene frequencies (BAZYKIN 1969; ENDLER 1977) and quantitative traits (BARTON 1999; LEIMAR *et al.* 2008), and other empirical studies mostly dedicated to the description of clinal patterns in organisms like *Drosophila* spp. (HALLAS *et al.* 2002; WEEKS *et al.* 2002), *Populus tremula* (INGVARSSON *et al.* 2006), *Quercus petrea* (ZANETTO and KREMER 1995), *Pinus sylvestris* (GARCIA-GIL *et al.* 2003), *Arabidopsis thaliana* (KRONHOLM *et al.* 2012), and yet other plant species (SAVOLAINEN *et al.* 2007). However, no attempt to measure selection in any of these or any other large-scale systems has been carried out, to our knowledge.

Still in the context of expanding populations, EDMONDS *et al.* (2004) proposed that the formation of (genotypic) allele-frequency clines across environmental gradients could also (and mainly) be caused by a purely neutral process during range expansion: the allele-surfing phenomenon, further studied and named by KLOPFSTEIN *et al.* (2006). In populations undergoing a range expansion, mutations arising at the front of the wave of expansion can “surf” on this wave and increase in frequency simply due to a series of founder effects. This surfing leaves behind a pronounced cline in allele frequencies, which may in turn have an effect on a phenotype, generating a phenotypic cline. Recent studies are bringing a growing body of evidence that allele surfing alone is capable of producing many of the allele-frequency clines observed in natural populations (CURRAT *et al.* 2006; EXCOFFIER and RAY 2008; HOFER *et al.* 2009). A recent finding even shows that range expansions might allow for the accumulation of deleterious mutations generating an 'expansion load' in populations of recently colonized areas (PEISCHL *et al.* 2013).

However, there is also evidence that adaptive processes may occur during range expansions, bringing about the idea of adaptive clines. For example, WHITE *et al.* (2013) recently found indications of adaptive evolution in an ongoing range expansion in Irish bank voles, where several genes related to immune and behavioral systems were shown to form consistent clines across three independent transects of the expansion. Also, it appears that dispersal ability itself is a trait commonly affected by selection in range expansions: higher dispersal is often selected for in the margins of an expansion, as theoretical analyses suggest (TRAVIS and DYTHAM 2002). Empirical support for this finding has been documented in several species (HUGHES *et al.* 2007; MONTY and MAHY 2010; MOREAU *et al.* 2011). Furthermore, rapid adaptation to climate variation also facilitates range expansion, as has been verified in the invasive plant *Lythrum salicaria* in North America (COLAUTTI and BARRETT 2013). The body of evidence favoring selection in range-expansion systems is substantial, and it often includes the examples of the (continental) large-scale clines mentioned above, as well (BAZYKIN 1969; ENDLER 1973; ENDLER 1977; BARTON 1999; LEIMAR *et al.* 2008). One particularly interesting case in large-scale cline and range expansion systems is the European barn owl (*Tyto alba*) and its coat-color cline (ANTONIAZZA *et al.* 2010). In this species, a gradient of colors has established across Europe, probably during or after a post-glacial range expansion, with white morphs nearly fixed in the southwest and dark-brown morphs in the northeast. This and the above-mentioned cases all suggest selection has been acting. However, the current challenge persists in (i) distinguishing neutrality from selection and (ii) properly measuring the strength of natural selection in large-scale clinal systems involved in range expansions.

Here, we take advantage of spatially explicit simulations to investigate the role of selection in the context of range expansions. First, we assess the ability of selection to leave a distinctive signature of its activity on the populations, despite the occurrence of the complicating demographic effects of range expansions (e.g. allele surfing). Second, with approximate Bayesian computation (ABC) (BEAUMONT *et al.* 2002), we address the detection and estimation of natural selection operating in this system. Finally, focused on the estimation of selection, we also explore the effect of other demographic and genetic parameters (nuisance parameters) on the accuracy of the selection estimates. Variations in these parameters may affect the probability of allele surfing. Therefore assessing the robustness of selection estimates across these parameters can bring valuable insight on the interplay between neutrality and natural selection in the ubiquitous demographic scenario of range expansions.

Material & Methods

Range expansion – Simulations were run in a rectangular world 5 patches wide and 51 (0-50) patches long (Fig. 1A) in the program quantiNEMO (NEUENSCHWANDER *et al.* 2008a). To mimic a range expansion, only the left-most patches started the simulations occupied at their carrying capacity (K = 100). These five patches evolved without any range expansion for arbitrary 100 generations in order to establish a background of genetic diversity, mimicking a refugium. The colonization of the remaining patches occurred after this initial phase and lasted 400 generations, at a speed that depended on the migration rate (m, uniform [0.1, 0.4]) and the intrinsic growth rate of each patch (r, uniform [0.2, 0.8]). We further varied narrow-sense heritability value (h^2, details below) and mutation rate (μ, log-uniform [10^{-5}, 10^{-2}]), were used as nuisance parameters to test the robustness of the selection-related parameter's estimates. As neutral genetic markers, ten multi-allelic loci were simulated with the same mutation rate implemented for the quantitative loci (μ) and a single-step mutation model, mimicking microsatellite markers.

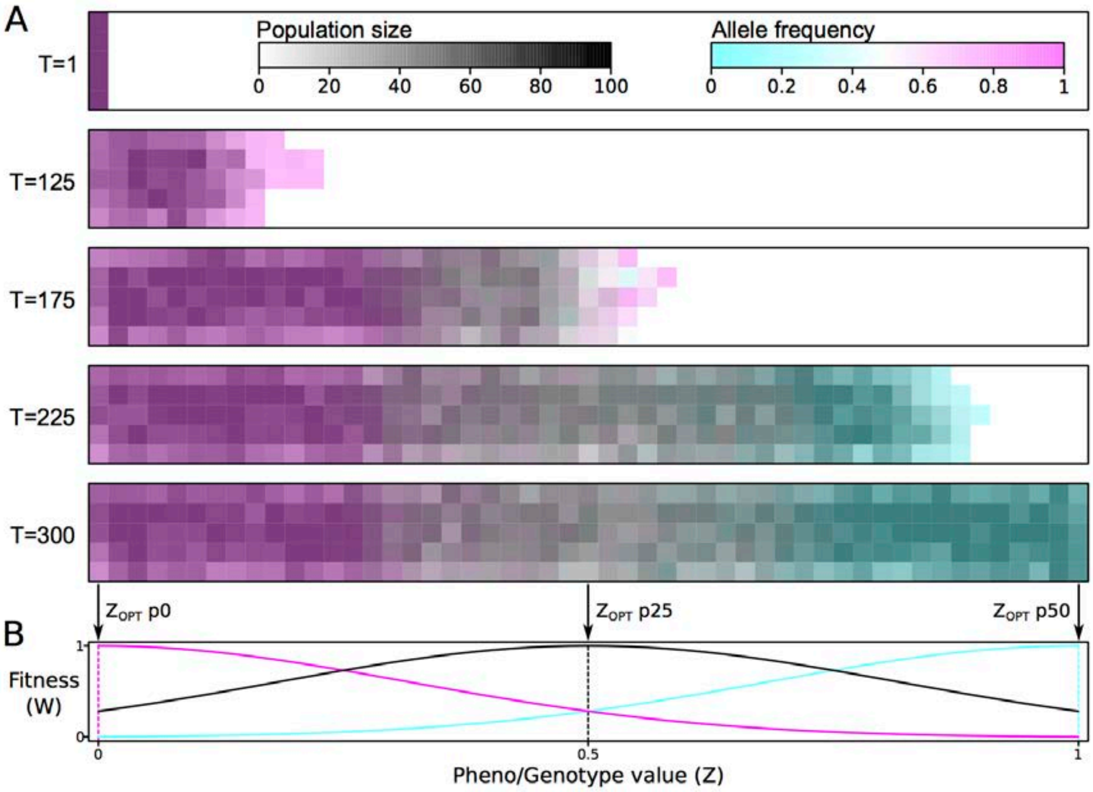

Figure 1: Implementation of the simulations with range expansion over a selection gradient. In **A**, the range expansion process over 300 generations (T), across the simulated map (51x5 demes). Two layers overlap here: population size (gray scale, underneath) and frequency of the allele adapted to the left-hand side of the map (cyan-magenta scale). In **B**, the fitness landscape for three patches from above (p0 magenta, p25 black, and p50 cyan) with selection intensity ω=0.1 and pheno/genotype space defined between 0 and 1. Note that the x-axis in B (Z-value) is different from the one in A (deme position p).

Selection implementation – Fig. 1B illustrates how selection was implemented: we assumed a local hard stabilizing-selection scheme with a gradient of optima along the colonization path. On the left-hand side of the map, the selective optimum was defined at one extreme of the phenotypic range ($Z_{OPT} = 0$); while, at the right-hand side, it was set to the other extreme ($Z_{OPT} = 1$). Each patch along the colonization path had a different optimum value (Z_{OPT}), linearly distributed between 0 and 1. Individual fitness is given by the function:

$$W_{ij} = e^{-\frac{\left(Z_{ij}-Z_{OPTj}\right)^2}{2\omega^2}}$$

where W_{ij} is the fitness of individual i from patch j with phenotype Z_{ij}, where the patch optimum is Z_{OPTj} and the selection intensity (identical for all patches) is given by ω. This latter parameter determines the strength of selection in our model (ω, log-uniform [0.1, 100], Fig. 2A). The ω parameter translates directly into a selection coefficient (s) (Fig. 2B) according to equation:

$$s = 1 - e^{-\frac{1}{2\omega^2}}$$

where s is the selection coefficient – defined as the difference in fitness between the two extreme pheno/genotypes ($Z = 0$ or 1) at any of the ends of the map – and ω is selection intensity, as already defined above. Part of the phenotype is environmentally determined, depending on trait heritability (h^2). We explored a wide range of heritability values (h^2, uniform [0.01, 1]), kept constant over time within the same simulation. Our goal is to estimate the selection coefficient (s), having nuisance parameters corresponding to the heritability of the trait (h^2), migration (m), mutation (μ) and growth (r) rates.

Six genetic architectures – Six different genetic architectures were implemented for the trait under selection where the allelic effects were entirely additive within and between loci. First, we assumed a trait encoded by one locus and two co-dominant alleles (1L2A). In this case, only one mutation was needed to make the leap between the two extremes of phenotype. The second model still involved only one locus, but with multiple alleles (1L10A), whose effects on the phenotype were linear and additive. Here, there are only two alleles completely adapted to the two extremes of the environmental gradient; all other alleles have intermediate values, which are able to match the intermediate optima along the colonization range. The third genetic architecture was that of a trait encoded by ten bi-allelic loci (10L2A) where all loci are required to adapt to obtain the extreme phenotypes. A second version of this architecture was one with the same number of loci and alleles, but with allelic effects large enough for a mutation at a single locus to allow for perfect adaptation to the extremes (10L2A+). A fifth architecture involved 10 alleles at 10 loci (10L10A), similar to 1L10A, but extended to ten independent loci. Similar to the extension of large allele effects applied in 10L2A+, a sixth architecture was defined with the possibility of any given locus as being able to modify the phenotype across its complete range (10L10A+). Mutation rate was scaled to the number of loci encoding the trait, so that the trait's mutation rate was the same across architectures (i.e. it was 10× lower for each locus in the 10L architectures).

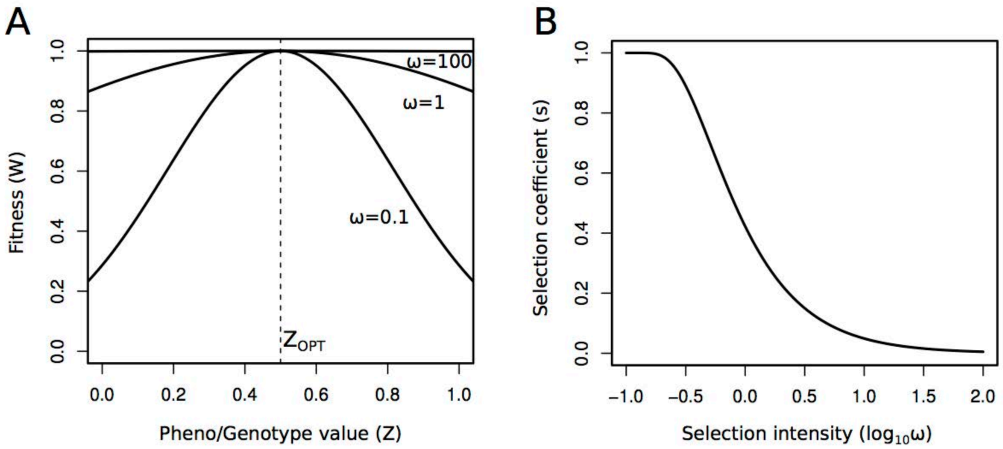

Figure 2: Fitness distribution and selection coefficient under different selection intensities (ω). In **A**, different fitness distributions with Z_{OPT} always at 0.5, as in patch p25 (see Fig. 1B), depicting the extremes of the ω prior distribution ω=0.1 and 100. In **B**, the effect of ω on the difference of fitness [i.e. selection coefficient (s)] between opposing pheno/genotype values at the extreme patches (p0 and p50).

ABC for selection – One suitable way to address complex evolutionary question is to implement approximate Bayesian computation (ABC). With this approach, one can assess the probability of different models and parameter values therein via summary statistics, thus dismissing the need of an exact likelihood function (BEAUMONT *et al.* 2002). Summary-statistic values are taken from the observation (i.e. the real populations) and compared to the values of the same statistics obtained in simulations. A large number of simulations are then used to explore different combinations of parameter values; the simulations that better match the summary statistics values of the observation are then used to draw a posterior distribution of parameter values. As a Bayesian method, ABC can (and should) incorporate prior information on the parameter distributions into the simulated model. Here, we applied ABC to the estimation of selection in a spatially explicit setting involving range expansions. Since this a simulation study, the observations were also taken from the simulations in the form of pseudo-observations (see below).

ABC: summary statistics – Based on our previous experience with a similar set-up in natural populations of barn owls (ANTONIAZZA *et al.* 2010; ANTONIAZZA *et al.* in prep.), we decided to focus on isolation-by-distance (IBD) pattern statistics as the statistics more likely to reveal the presence of selection: From the correlation between pairwise geographic distance and pairwise pheno/genotypic distance, we extracted the mean, slope and sum of residuals for ten neutral multi-allelic markers (F_{ST}), and the phenotype (Q_{ST}). Finally, we also retained the difference of slopes of IBD between the phenotype and the neutral markers (Δ-slope), which represents how much steeper is the differentiation in the quantitative trait when compared to the one produced by the neutral loci (Fig. S1).

ABC: parameter estimates and estimability assessment – We tested the precision and accuracy of parameter estimates through ABC's validation approach as implemented in ABCtoolbox (WEGMANN *et al.* 2010). Since the actual parameter values for all simulations are known (pseudo-observations), the ABC parameter-estimation pipeline was used to assess the quality of the estimates (i.e. how close the estimates were to the actual values). This was done by comparing 1000 of these estimates with their actual pseudo-observed values taken directly from the simulations, for each one of the genetic-architecture models. This procedure involved retaining the 1000 (out of ~1 million) simulations with summary statistics values closest to the pseudo- observation's, and then to use locally weighted linear regressions to obtain the posterior distributions for the parameter estimates (WEGMANN *et al.* 2010). The overall estimability of selection coefficient for the different architectures was assessed using the coefficient of determination (R^2) of the regression between the true value of the parameter (pseudo-observation) and the parameter point estimate (given by the mode of the posterior distribution) (NEUENSCHWANDER *et al.* 2008b). Two other statistics were also used to assess estimability: the root mean square error (RMSE), which depicts the prediction errors of our model by means of the mean absolute differences between pseudo-observations and estimates (WEGMANN and EXCOFFIER 2010); and proportion of the estimated posterior encompassing the pseudo-observed value for 50% and 95% of the higher-posterior density intervals (proportion of HPD50% and 95%). This latter statistics may indicate a low accuracy, when proportion of HPD50% << 0.5, or HPD95% << 0.95; or excessive conservativeness, when proportion of HPD50% >> 0.5, or HPD95% >> 0.95. Ideally, HPD50% and 95% should be exactly 0.5 and 0.95, respectively.

Moreover, to assess the effect of the nuisance parameters (m, r, μ, h^2) on the estimability of selection coefficients, a second test was devised in which the parameter space of each one of the nuisance parameters was restricted to ten quantiles. The estimations of selection coefficient were obtained only in that restricted space. For example, heritability (h^2) varied randomly from 0.01 to 1 across all simulations. To test whether estimates of selection were robust to a predetermined h^2 value, we separated the simulations in ten different sets according to different quantile intervals of the h^2 prior distribution – e.g. the first interval includes the simulation in which h^2 ranges from 0.01 to ~0.1. For each of the h^2 intervals, we obtain estimations of selection coefficient (s) that were then compared to their pseudo-observed value. This was also done for the other three nuisance parameters (m, r and μ) and across all six genetic architectures. Quantiles of the parameter values, instead of fixed bins, had to be used in order to insure that all estimates were made based on the same number of simulations. This is because the inherent sampling process, combined with the failure of some simulations (see supplement), does not necessarily leads to the same density of simulations across the whole parameter space. So, for each quantile interval, 1000 estimates were run with 500 retained simulations, and the estimability was again measured by means of R^2, allowing for comparisons across the quantiles.

Results

Overall, the statistics related to the regressions between pairwise differentiation (Q_{ST}) and pairwise geographical distances were very sensitive to variation in selection strength, regardless of the genetic architecture implemented (Fig. 3). In particular, the difference of Q_{ST} and F_{ST} IBD slopes (Δ-slope) showed to be particularly responsive to small selection coefficients, while mean differentiation on the phenotype (mean Q_{ST}) was more sensitive to moderate and high selection coefficients. Additionally, as expected for independent neutral loci, the statistics related to F_{ST} alone did not vary with the selection coefficients (results not shown). For nearly all architectures, mean Q_{ST} showed a constant quasi-linear increase with higher selection coefficients (Fig. 3A). The only two exceptions were the 1L2A and the 10L2A+ (with large-effect alleles) architecture models. In fact, these two architectures showed very concordant responses also in the other statistics, such as Δ-slope (Fig. 3B). In both cases, one can observe a lack of points for high selection coefficient values ($s > 0.5$). Indeed, these simulations failed to colonize the entire habitat (further examined below in 'Discussion'). Moreover, Δ-slope, for all architectures, reaches an asymptote when $s > 0.4$. This is because, when selection is very strong, even closely neighboring demes are highly differentiated (high Q_{ST}). This leads to high mean Q_{ST}, but limits (or even reduces) the values obtained for the slope of differentiation across the environmental gradient (Fig 3B). Noteworthy are also the similarities between 1L10A and 10L10A+.

The quality of estimates for selection coefficient (s) in all models was high (Table 1, Fig. 4). The genetic-architecture models 1L10A, 10L2A, 10L10A and 10L10A+ had particularly high coefficients of determinations ($R^2 > 0.9$), with 1L2A and 10L2A+ falling shortly behind ($R^2 > 0.7$). This difference among the architectures derives from the differences in the summary statistics (above), where simulations with $s > 0.5$ failed to leave any signature on the summary statistics (Fig. 3), resulting in a limited range of s values (Fig. 4). Furthermore, the root mean square error values were proportionally low for all architectures (RMSE ≈ 5 to 9% of s estimates), implying a very high accuracy of estimates. The proportion of posterior-estimate distributions that encompassed the pseudo-observed value – both with HPD50% and 95% – resulted in conservative estimates (Table 1), with proportion values always larger than the HPD interval. This suggests that, even though accurate, the posterior distributions are not necessarily very precise, with rather wide ranges.

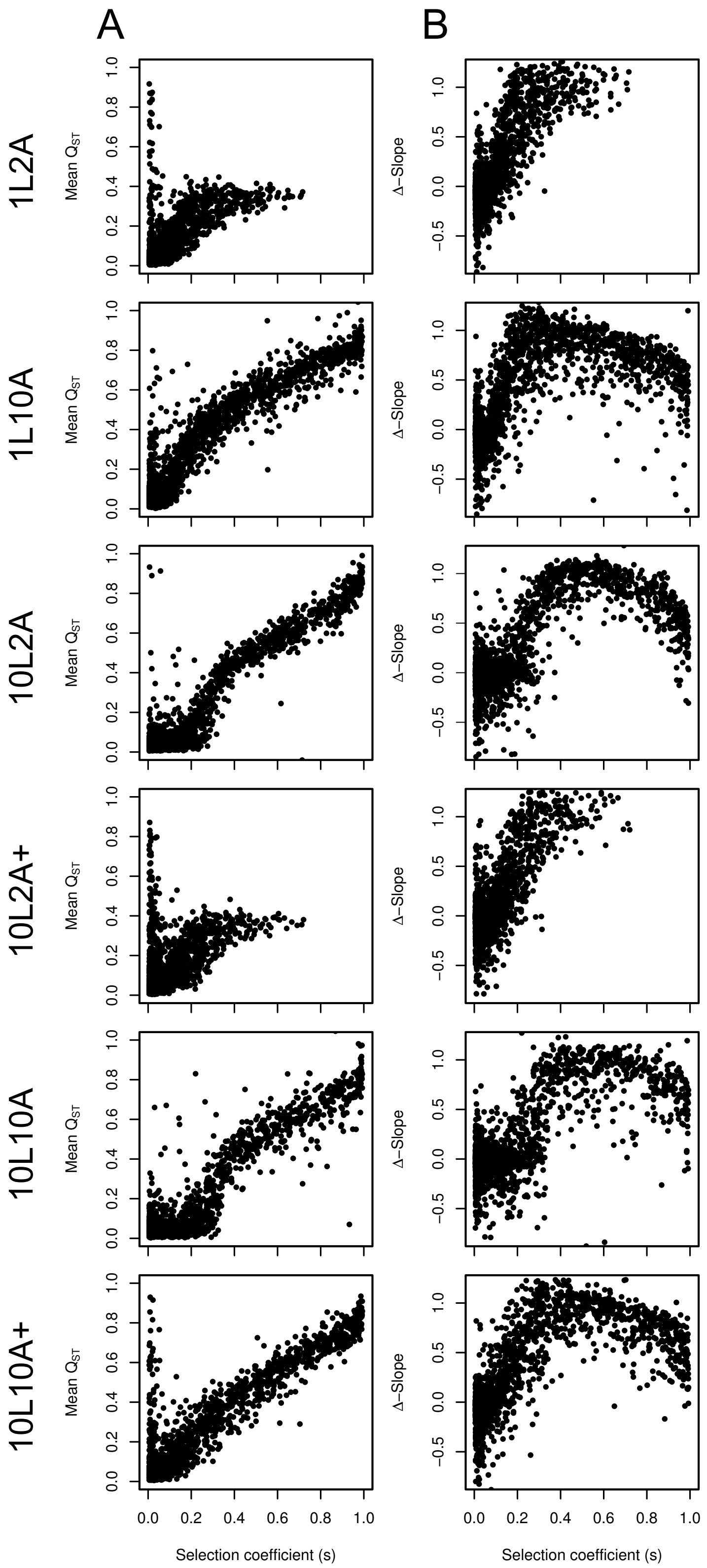

Figure 3: The relation between selection coefficient (s) and the most informative pattern statistics used to assess the selection coefficient. For all six architectures, in **A**, the response of mean differentiation across populations (Mean Q_{ST}); and in **B**, the response of the difference between the Q_{ST} and the neutral F_{ST} slopes of IBD (Δ-Slope).

Remarkably, in our simulations, the estimability results are robust to the variation in the nuisance parameters and to the position in the largest part of the nuisance parameters' space, with the clear exception of lower values of heritability ($h^2 < 0.1$) for all architectures and also, to a lesser extent, lower values of mutation rate for some architecture models (Fig. 5). Interestingly, variation in migration (m) and growth rate (r) in the interval explored (m = [0.1, 0.4] and r = [0.2, 0.8]) has very little effect. Here too, there seems to be a ranking of estimation quality among the genetic-architecture models across the nuisance parameter quantiles: 1L2A and 10L2A+ being the worse (but still good); followed by 10L10A; and then having 10L10A+, 10L2A and 1L10A as the better ones.

Table 1: Assessment of selection coefficient (s) estimability for all genetic architectures. R^2 stands for the coefficient of determination of the pseudo-observation on the estimates; RMSE is root mean square error of the estimates; and Prop. HPD50% and HPD95% represent the proportion of posterior distributions encompassing the pseudo-observed value. These values were obtained based on 1000 estimates, with 1000 retained simulations out of 1 million simulations, under a stabilizing hard selection system.

Architecture	R^2	RMSE	Prop. HPD50%	Prop. HDP95%
1L2A	0.837	0.049	0.726	0.988
1L10A	0.958	0.065	0.646	0.982
10L2A	0.952	0.066	0.703	0.989
10L2A+	0.738	0.056	0.665	0.992
10L10A	0.911	0.087	0.654	0.988
10L10A+	0.963	0.060	0.590	0.971

Discussion

We have shown that it is possible to estimate selection and its intensity in range expansions by taking advantage of the information contained in IBD-derived statistics and by using spatially explicit simulations. Even though plenty of variance in the response of the summary statistics was observed when comparing the different genetic-architecture models, in all cases, selection left a distinctive signature on these statistics. It seems, however, that the probability of the populations to respond to selection was not the same across all architectures. In a nutshell, the more alleles and loci encoded the trait; the better was the estimation of the selection coefficient.

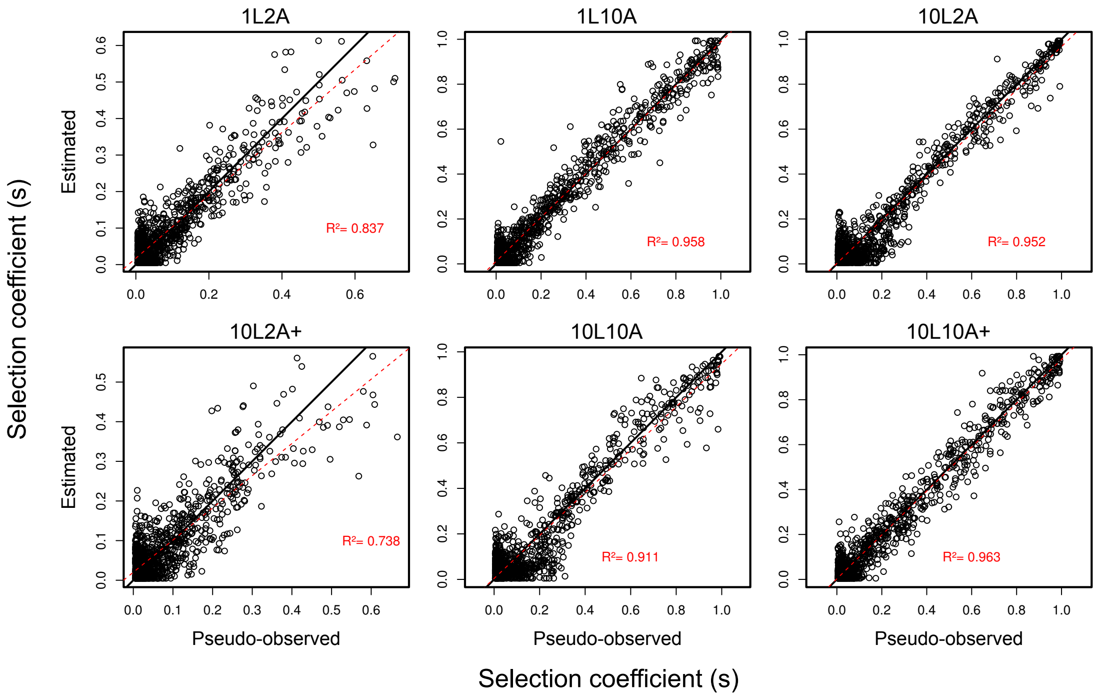

Figure 4: Validation plots, pseudo-observed vs. estimated, for selection coefficient (s). For each genetic-architecture model, a plot of 1000 simulations' actual selection coefficient values (s) against their estimates (open circles). The back line stands for the perfect diagonal; and the red dashed line, the calculated linear regression. Coeffiecients of determination of the pseudo-observation on the estimates (R^2) are also reported in red.

The architectures can be divided in three groups: (i) 1L10A and 10L10A+ with very high R^2 and low RMSE (i.e. very good estimability), (ii) 10L2A and 10L10A with still high R^2 and low RMSE values but with a distinct signature in the Δ-slope statistic (Fig. 3B), and (iii) 1L2A and 10L2A+ with slightly worse R^2 and RMSE results. Not surprisingly, these last two architectures are also the ones that present the least number of allele combinations (within the phenotypic range between Z = 0 and 1) that could lead to adaptation across the selection gradient. The 1L2A model has only three possible genotypes to be translated into phenotypes. In essence, this architecture is just as capable as the others to adapt to the two extremes and the exact center of our simulated environment (patches p0, p50 and p25, respectively). However, this does not apply for any of the patches in between. In these other patches, there is no combination of alleles that would make an individual perfectly adapted to the local conditions. This same explanation applies to 10L2A+ because the large-effect alleles turn up to make too big a leap in between pheno/genotypic values (Z in Fig. 1). Indeed, if a second locus mutates as well in 10L2A+, the Z-value of the resulting phenotype would almost certainly fall outside the range of adapted phenotypes in all patches (Z = 0 to 1). This is why, when selection is too strong ($s > 0.5$), simulations failed to finish the colonization due to the recurrent extinction of pioneer populations. Conversely, all the other architectures present many more Z-value combinations allowing to locally adapt to all patches across the colonization range. These results may suggest that adaptation may be easier to occur when many loci and alleles contribute to a trait – offering several to many combinations of loci and alleles in order to adapt to the local conditions – in agreement with previous studies (LE CORRE and KREMER 2012).

It is important to highlight the impact of the inclusion of spatial information in the understanding of the effect of selection in range-expansion scenarios. The process of range expansion is essentially a spatial phenomenon and, to fully understand its outcome, a spatially explicit approach is warranted. Even though some of the statistics we used – mean F_{ST} and Q_{ST} – do not explicitly contain spatial information, it was only with the addition of Δ-slope and the other IBD-associated statistics that we managed to grasp the full extent of the of the effect of selection in range-expansion processes. The importance of the spatial dimension in population genetics is not a novel idea, though. It has been explored in numerous previous publications, both in the disciplines of phylogeography (AVISE *et al.* 1987; DINIZ-FILHO *et al.* 2008) and, more recently, in landscape genetics (MANEL *et al.* 2003). Studies looking for signatures of selection, however, have been systematically neglecting the relevance of the spatial distribution of genes and phenotypes (LI *et al.* 2012).

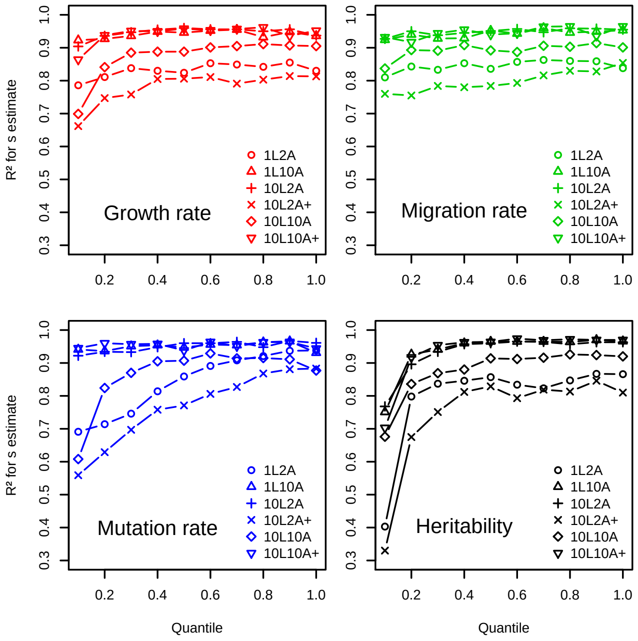

Figure 5: Estimability assessment across the nuisance-parameter space, for all genetic architectures. In each panel, the estimability of selection coefficient (by means of R^2) is shown for ten different quantiles of the realized prior distributions fo the four nuisance parameters (each panel) and all six genetic architectures (within panels).

Furthermore, combining more than one pattern statistics (at least mean Q_{ST} and Δ-slope, Fig. 3) seems to be of key importance to properly assess the effect of selection on populations facing range expansions. For instance, the analysis of mean Q_{ST} alone could lead to false positives when selection is very low (virtually zero), given that a few observations of high overall differentiation appear in these quasi-neutral conditions

(Fig. 3A). Also, looking at Δ-slope alone could lead to false negatives – or simply lack of information – when selection is too strong, leading to less steep slopes than the ones observed at intermediate selection coefficients (Fig. 3B). Therefore, to properly benefit from our proposed ABC approach, we believe that one should always, of course, consider all available information contained in the different IBD pattern statistics.

Even though we modeled selection via intensity of selection (ω) – a parameter widely used in quantitative genetics (FALCONER and MACKAY 1996) – we decided to estimate selection through selection coefficient (s), which is a more common measure in population genetics (HARTL and CLARK 2007). Selection coefficient is a parameter whose effect on fitness (W) is directly accessible ($W = 1 - s$), making biological interpretation easier. Also, while ω had to be treated in the logarithmic scale (to obtain a more linear relation with the summary statistics), s could be dealt with in a linear scale. Besides, the results for estimability calculated for $\log_{10}\omega$ showed only a slight trend to lower R^2 values and did not differ substantially from the ones obtained with s (Table S1). Regarding the scale of selection coefficient here, it is worth to remember that it concerns the difference in fitness in the extreme patches and the difference in fitness between the extreme pheno/genotypes (p0 and p50, Fig. 1). It becomes smaller as one approaches the center of the map and/or compares closer pheno/genotypes, and therefore represents the maximum strength of selection operating in the system.

We mentioned that some simulations "failed to finish the colonization altogether". This requires further explanation. By failed simulations, we do not necessarily mean simulation where the population went extinct, but actually simulations that resulted in missing-data (NA) for any of the statistics. First, the simulations were run assuming a hard-selection system (i.e. individual fitness is absolute). So – even though local populations could react to loss of individuals via population growth (r) – if selection was too strong and no locally-adapted individuals were yet present at the population, that specific deme would go extinct delaying or stopping the wave of expansion. Alternatively, we also ran the same simulations with a soft-selection system (supplementary material). These showed a lower failure rate, but did not affect further results, suggesting that the approach presented here is also robust to the softness of the selection implemented. Second, some architecture models lead to higher failure rates than others, predominantly due to the non-colonization effect described above. This is again related to the limited combinations of loci and alleles observed in architectures 1L2A and 10L2A+. As a result, the realized prior distribution (i.e. the parameter distribution after the removal of simulations containing NAs) for selection intensity (ω) – and therefore selection coefficient (s, as in Fig. 2) – was altered for these two architectures, being limited to $\omega = 10^{-0.5}$ to 10^2 ($s \approx 0.8$ to 0, respectively, Fig. S2 and S3). Beyond selection strength, for the other simulation parameters (i.e. nuisance parameters), there was no differential effect of the architecture model on the way these parameters produced simulations containing missing data. There was, however, a more elevated missing data production, for all architectures, associated with low mutation rates (when $\mu < 10^{-4}$), when not enough variation was produced to adapt to new environments; low growth rates ($r < 0.3$), when the negative effect of higher selection coefficients was stronger on the populations; and, to a lesser extent, higher migration rates, where the homogenizing effect of migration more often erased the differentiation signatures created by selection. As a result, the prior distributions for the nuisance parameters were altered after the removal of such failed simulations (Fig. S2). Consequently, the ten quantiles presented in Fig. 5 do not necessarily represent 10% intervals of the original prior distributions, but rather regular intervals taken from realized prior distributions. The analysis was done this way in order to have the same number of simulations out of which to make the estimates in each interval, allowing for a balanced comparison of estimability across quantiles.

The estimability of selection was little affected by variation in the nuisance parameters, as R^2 remained well above 0.7 for all genetic architecture models across most of these parameters' distributions. Some of the architectures seemed to be more sensitive to the noise caused by these parameters than others: Again, 1L2A and 10L2A+ showed to be the most sensitive models, probably, due to the lack of possible genotypic combinations, limiting adaptation to intermediary positions across the environmental gradient, as discussed above. However, the variation in mutation rate also had some effect on these architectures. The lower the mutation rate, the harder to deal with very strong selection, especially when combinations are limited. Another architecture in which selection estimability strongly responded to mutation rate was 10L10A. Curiously, this is the one with highest number of genotype combinations. This can be explained by the fact that it also is the architecture that needs the most mutations in order to adapt to the opposite environmental conditions during the range expansion. All ten loci need to adapt by fixing one of ten possible alleles each. Finally, as one could already expect, low values of heritability led to lower estimability for all architectures. Clearly, if the trait under selection has a very small genetic component, selection can do very little to affect the differentiation of the quantitative trait, leaving no signature of adaptation in the pattern statistics we explored, or any other statistics one could think of, as well.

It is still computationally expensive to run the individual-based spatially explicit simulations required to study the evolution of quantitative traits in range expansions, especially with several models of genetic architecture (e.g. ~350 CPU days for 1 million simulations on a Linux server with a 2.4GHz Intel Xeon processor). This is because an

ABC implementation generally requires many simulations (at least 1 million) to obtain reliable parameter estimates (FAGUNDES *et al.* 2007; NEUENSCHWANDER *et al.* 2008b), even though this can dependent at a large extent on the number of the parameters to be estimated (i.e. the dimensions of the parameter space to explore). Alternatively, improvements on the ABC algorithm such as MCMC-ABC (WEGMANN *et al.* 2009) can help reducing the number of simulations needed for investigating a given question. Besides, selection was not the only parameter varying in our model. Nuisance parameters, even though not estimated, also affect the parameter space to be explored by the simulations. These do not have to be used, though: We added them to our analysis to assess the robustness of our estimates, but this does not need to be done in empirical studies. An approach that could be followed in such studies would be a two-step ABC analysis (BAZIN *et al.* 2010), where (i) one would determine a neutral demographic background based on neutral markers and coalescent simulations, and (ii) then use the estimates of this previous step as priors for the following one in which individual-based simulations would be run to explore a different set of fewer parameters (e.g. selection coefficient and heritability), assuming that the effects of selection on demography would have already been captured in the first step.

Contrary to an impression one might get reading the recent theoretical literature on range expansions (KLOPFSTEIN *et al.* 2006; TRAVIS *et al.* 2007; EXCOFFIER *et al.* 2009; PEISCHL *et al.* 2013), selection is able to operate in such scenarios. Recent empirical studies have been showing evidence that adaptation has occurred in several cases (HUGHES *et al.* 2007; ANTONIAZZA *et al.* 2010; MONTY and MAHY 2010; BUCKLEY *et al.* 2012; ANTONIAZZA *et al.* in prep.). When compared to allele surfing, selection seems to be much more efficient in producing differentiation across the range of an expansion, according to our results. Even though we observed consistent isolation by distance in the neutral loci (proxy for pure allele surfing), this isolation was always much lower than what was observed for the trait under selection.

The direct observation of some simulations provided evidence that locally maladapted variants could appear and reach relatively high frequencies during the range expansion process (Fig. S4), but these events tended to be transient and were quickly erased by selection, leaving virtually no signature after the whole map had been occupied. This observation may be the result of the model implemented here, where only one locus or a few loci were involved with selection and, therefore, could bear locally maladaptive (deleterious) variants. Another theoretical study, focused on the evolution of genetic load, provided evidence that, when many loci are involved, the overall deleterious load of populations undergoing range expansions tends to increase (PEISCHL *et al.* 2013). Indeed, there seems to be a decrease in the efficiency of purifying selection in purging a genome-wide deleterious load during range expansion (i.e. expansion load). However, here we investigated a process involving positive selection acting on one specific phenotypic trait whose genetic architecture was relatively simple. It is in this situation, we showed that natural selection during range expansions is still effective. Furthermore, in real populations, the simultaneous occurrence of adaptation at a given trait with the accumulation of an expansion load is perfectly possible and may be one explanation for the success of so many range expansions observed in nature. The combined effect of these two processes, however, remains to be more carefully investigated in the future.

Even though neutrality (including background selection) (KIMURA 1984) should always be the null hypothesis for any investigation of a process leading to a given observed pattern, we believe that here we have gathered sufficient *in silico* evidence that selection can operate on range expansion scenarios, leaving a distinguishable signature in spatially explicit statistics. Furthermore, this signature allows estimating the strength of selection operating on the study system and could be promptly used in empirical studies investigating selection in range expansion scenarios – which could be post-glacial recolonizations, species invading new habitats, or populations coping with environmental changes. All of these processes were and still are very common, not only in temperate regions (HEWITT 2004), but also anywhere else on the globe, rendering the spatially explicit ABC approach presented here particularly valuable.

Acknowledgments

The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. We would like to thank Laurélène Faye and Rémi Matthey-Doret for helping in initial stages of this project and Sylvain Antoniazza for many fruitful discussions that helped in the development of this study. JG had financial support from the Swiss National Science Foundation (SNSF) grant number XXXXXX.

References

ANTONIAZZA, S., R. BURRI, L. FUMAGALLI, J. GOUDET and A. ROULIN, 2010 Local adaptation maintains clinal variation in melanin-based coloration of European barn owls (Tyto alba). Evolution **64:** 1944-1954.

ANTONIAZZA, S., R. KANITZ, S. NEUENSCHWANDER, R. BURRI, A. GAIGHER *et al.*, in prep. Natural selection in a post-glacial range expansion: the case of the colour cline in the European barn owl.

AVISE, J. C., J. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB *et al.*, 1987 Intraspecific Phylogeography - the Mitochondrial-DNA Bridge between Population-Genetics and Systematics. Annual Review of Ecology and Systematics **18:** 489-522.

BARTON, N. H., 1999 Clines in polygenic traits. Genet Res **74:** 223-236.

BARTON, N. H., and G. M. HEWITT, 1985 Analysis of Hybrid Zones. Annual Review of Ecology and Systematics **16:** 113-148.

BAZIN, E., K. J. DAWSON and M. A. BEAUMONT, 2010 Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. Genetics **185:** 587-602.

BAZYKIN, A. D., 1969 Hypothetical mechanism of speciation. Evolution **23:** 685-687.

BEAUMONT, M. A., W. ZHANG and D. J. BALDING, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025-2035.

BUCKLEY, J., R. K. BUTLIN and J. R. BRIDLE, 2012 Evidence for evolutionary change associated with the recent range expansion of the British butterfly, Aricia agestis, in response to climate change. Mol Ecol **21:** 267-280.

COLAUTTI, R. I., and S. C. BARRETT, 2013 Rapid adaptation to climate facilitates range expansion of an invasive plant. Science **342:** 364-366.

COLINVAUX, P. A., P. E. DE OLIVEIRA and M. B. BUSH, 2000 Amazonian and neotropical plant communities on glacial time-scales: The failure of the aridity and refuge hypotheses. Quaternary Science Reviews **19:** 141-169.

CURRAT, M., L. EXCOFFIER, W. MADDISON, S. P. OTTO, N. RAY *et al.*, 2006 Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens" and "microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". Science **313:** -.

DINIZ-FILHO, J. A. F., M. P. DE CAMPOS TELLES, S. L. BONATTO, E. EIZIRIK, T. R. O. DE

FREITAS *et al.*, 2008 Mapping the evolutionary twilight zone: molecular markers, populations and geography. Journal of Biogeography **35:** 753-763.

EDMONDS, C. A., A. S. LILLIE and L. L. CAVALLI-SFORZA, 2004 Mutations arising in the wave front of an expanding population. Proc Natl Acad Sci U S A **101:** 975-979.

ENDLER, J. A., 1973 Gene flow and population differentiation. Science **179:** 243-250.

ENDLER, J. A., 1977 *Geographic Variation, Speciation, and Clines*. Princeton University Press, Princeton.

EWENS, W., 1977 Population Genetics Theory in Relation to the Neutralist-Selectionist Controversy, pp. 67-134 in *Advances in Human Genetics 8*, edited by H. HARRIS and K. HIRSCHHORN. Springer US.

EXCOFFIER, L., M. FOLL and R. J. PETIT, 2009 Genetic Consequences of Range Expansions. Annual Review of Ecology, Evolution, and Systematics **40:** 481-501.

EXCOFFIER, L., and N. RAY, 2008 Surfing during population expansions promotes genetic revolutions and structuration. Trends in Ecology & Evolution **23:** 347-351.

FAGUNDES, N. J., N. RAY, M. BEAUMONT, S. NEUENSCHWANDER, F. M. SALZANO *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci U S A **104:** 17614-17619.

FALCONER, D. S., and T. F. C. MACKAY, 1996 *Introduction to quantitative genetics*. Prentice Hall, Harlow, UK.

GARCIA-GIL, M. R., M. MIKKONEN and O. SAVOLAINEN, 2003 Nucleotide diversity at two phytochrome loci along a latitudinal cline in Pinus sylvestris. Molecular Ecology **12:** 1195-1206.

HALLAS, R., M. SCHIFFER and A. A. HOFFMANN, 2002 Clinal variation in Drosophila serrata for stress resistance and body size. Genetics Research **79**.

HARTL, D. L., and A. G. CLARK, 2007 *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA, USA.

HEWITT, G., 2000 The genetic legacy of the Quaternary ice ages. Nature **405:** 907-913.

HEWITT, G. M., 1996 Some genetic consequences of ice ages, and their role in divergence and speciation. Biological Journal of the Linnean Society **58:** 247-276.

HEWITT, G. M., 2004 Genetic consequences of climatic oscillations in the Quaternary. Philos Trans R Soc Lond B Biol Sci **359:** 183-195; discussion 195.

HEY, J., 1999 The neutralist, the fly and the selectionist. Trends in Ecology & Evolution **14:** 35-38.

HOFER, T., N. RAY, D. WEGMANN and L. EXCOFFIER, 2009 Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet **73:** 95-108.

HUGHES, C. L., C. DYTHAM and J. K. HILL, 2007 Modelling and analysing evolution of dispersal in populations at expanding range boundaries. Ecological Entomology **32:** 437-445.

INGVARSSON, P. K., M. V. GARCIA, D. HALL, V. LUQUEZ and S. JANSSON, 2006 Clinal variation in phyB2, a candidate gene for day-length-induced growth cessation and bud set, across a latitudinal gradient in European aspen (Populus tremula). Genetics **172:** 1845-1853.

JEPSEN, J. U., S. B. HAGEN, R. A. IMS and N. G. YOCCOZ, 2008 Climate change and outbreaks of the geometrids Operophtera brumata and Epirrita autumnata in subarctic birch forest: evidence of a recent outbreak range expansion. J Anim Ecol **77:** 257-264.

KIMURA, M., 1984 *The neutral theory of molecular evolution*. Cambridge University Press.

KLOPFSTEIN, S., M. CURRAT and L. EXCOFFIER, 2006 The fate of mutations surfing on the wave of a range expansion. Molecular Biology and Evolution **23:** 482-490.

KRONHOLM, I., F. X. PICO, C. ALONSO-BLANCO, J. GOUDET and J. DE MEAUX, 2012 Genetic Basis of Adaptation in Arabidopsis Thaliana: Local Adaptation at the Seed Dormancy Qtl Dog1. Evolution **66:** 2287-2302.

LE CORRE, V., and A. KREMER, 2012 The genetic differentiation at quantitative trait loci under local adaptation. Mol Ecol **21:** 1548-1566.

LEIMAR, O., M. DOEBELI and U. DIECKMANN, 2008 Evolution of phenotypic clusters through competition and local adaptation along an environmental gradient. Evolution **62:** 807-822.

LI, J., H. LI, M. JAKOBSSON, S. LI, P. SJODIN *et al.*, 2012 Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol Ecol **21:** 28-44.

MANEL, S., M. K. SCHWARTZ, G. LUIKART and P. TABERLET, 2003 Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology & Evolution **18:** 189-197.

MONTY, A., and G. MAHY, 2010 Evolution of dispersal traits along an invasion route in the wind-dispersed Senecio inaequidens (Asteraceae). Oikos **119:** 1563-1570.

MOREAU, C., C. BHERER, H. VEZINA, M. JOMPHE, D. LABUDA *et al.*, 2011 Deep human genealogies reveal a selective advantage to be on an expanding wave front. Science **334:** 1148-1150.

MULLEN, L. M., and H. E. HOEKSTRA, 2008 Natural selection along an environmental gradient: a classic cline in mouse pigmentation. Evolution **62:** 1555-1570.

NEI, M., 2005 Selectionism and neutralism in molecular evolution. Mol Biol Evol **22:** 2318-2342.

NEI, M., Y. SUZUKI and M. NOZAWA, 2010 The neutral theory of molecular evolution in the genomic era. Annu Rev Genomics Hum Genet **11:** 265-289.

NEUENSCHWANDER, S., F. HOSPITAL, F. GUILLAUME and J. GOUDET, 2008a quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. Bioinformatics **24:** 1552-1553.

NEUENSCHWANDER, S., C. R. LARGIADER, N. RAY, M. CURRAT, P. VONLANTHEN *et al.*, 2008b Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. Mol Ecol **17:** 757-772.

PARMESAN, C., and G. YOHE, 2003 Nature A globally coherent fingerprint of climate change impacts across natural systems. Nature **421:** 37-42.

PEISCHL, S., I. DUPANLOUP, M. KIRKPATRICK and L. EXCOFFIER, 2013 On the accumulation of deleterious mutations during range expansions. Mol Ecol **22:** 5972-5982.

ROTH, D., B. HENRY, S. MAK, M. FRASER, M. TAYLOR *et al.*, 2010 West Nile virus range expansion into British Columbia. Emerg Infect Dis **16:** 1251-1258.

RUNDELL, R. J., and T. D. PRICE, 2009 Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. Trends Ecol Evol **24:** 394-399.

SAVOLAINEN, O., T. PYHÄJÄRVI and T. KNÜRR, 2007 Gene Flow and Local Adaptation in Trees. Annual Review of Ecology, Evolution, and Systematics **38:** 595-619.

THORPE, R. S., 1984 Primary and Secondary Transition Zones in Speciation and Population Differentiation - a Phylogenetic Analysis of Range Expansion. Evolution **38:** 233-243.

TRAVIS, J. M., and C. DYTHAM, 2002 Dispersal evolution during invasions. Evolutionary Ecology Research **4:** 1119-1129.

TRAVIS, J. M., T. MUNKEMULLER, O. J. BURTON, A. BEST, C. DYTHAM *et al.*, 2007 Deleterious mutations can surf to high densities on the wave front of an expanding population. Mol Biol Evol **24:** 2334-2343.

WAGNER, A., 2008 Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet **9:** 965-974.

WALTHER, G. R., A. ROQUES, P. E. HULME, M. T. SYKES, P. PYSEK *et al.*, 2009 Alien species in a warmer world: risks and opportunities. Trends Ecol Evol **24:** 686-693.

WEEKS, A. R., S. W. MCKECHNIE and A. A. HOFFMANN, 2002 Dissecting adaptive clinal variation: markers, inversions and size/stress associations in Drosophila melanogaster from a central field population. Ecology Letters **5:** 756-763.

WEGMANN, D., and L. EXCOFFIER, 2010 Bayesian inference of the demographic history of chimpanzees. Molecular Biology and Evolution **27:** 1425-1435.

WEGMANN, D., C. LEUENBERGER and L. EXCOFFIER, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics **182:** 1207-1218.

WEGMANN, D., C. LEUENBERGER, S. NEUENSCHWANDER and L. EXCOFFIER, 2010 ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics **11:** 116.

WHITE, T. A., S. E. PERKINS, G. HECKEL and J. B. SEARLE, 2013 Adaptive evolution during an ongoing range expansion: the invasive bank vole (Myodes glareolus) in Ireland. Mol Ecol **22:** 2971-2985.

ZANETTO, A., and A. KREMER, 1995 Geographical structure of gene diversity in Quercus petraea (Matt.) Liebl. I. Monolocus patterns of variation. Heredity **75:** 506-517.

Natural selection during range expansions: insights from a spatially explicit ABC approach

Ricardo Kanitz, Samuel Neuenschwander, Jérôme Goudet

Supplementary Material

Pattern statistics explained – Statistics related to patterns of isolation-by-distance (IBD) were used as summary statistics in our ABC analysis. From a scatterplot of geographic distance vs. genetic/phenotypic distance (F_{ST}/Q_{ST}), one can extract several statistics related to the spatial distribution of differentiation of neutral traits (F_{ST}) and quantitative traits (Q_{ST}). We used seven values to describe these relations: (1,2) mean F_{ST} and Q_{ST}, (3,4) slope of the regression of F_{ST} and Q_{ST} vs. geographic distance, the (5,6) logarithm of the sum of residuals for both F_{ST} and Q_{ST} regressions, and (7) the difference between the Q_{ST} and the F_{ST} slopes (Fig S1).

Soft selection model – As an addition to the hard selection model implemented in the main body of this study, we also tested the outcomes of a soft selection model, where the demes are refilled to their previous generation's population size readjusted by the population growth rate (r) and migration rate (m). So that, if selection is too strong, there is no local extinction caused by it. The results did not differ very much the ones obtained with the hard selection model, except for the range in which architectures 1L2A and 10L2A+ managed to respond to selection: here it covered the whole initially determined prior distribution, but a lower density in high selection coefficient values ($s > 0.5$, Fig. S2 and S3).

F_{STQ}-based estimates – We also ran estimates including information on the quantitative trait loci differentiation as part of the summary statistics (i.e. F_{STQ}). F_{STQ}

stands for F_{ST} applied to loci underlying the phenotype, as opposed to the phenotype itself (Q_{ST}). The resulting estimates were not significantly better than the ones already obtained with phenotype-based only statistics (Q_{ST}). Furthermore, considering empirical applications, most of the times the underlying loci of any given phenotype are unknown. Thus, avoiding the use of the F_{STQ}-related statistics is a more realistic approach.

Estimates for selection intensity (ω) – We also ran estimates for local stabilizing selection intensity (ω). This parameter was treated in the logarithmic scale in order to have a more linear relation with the summary statistics. The estimability results for the different models of genetic architecture (Table S1) are not significantly different from the estimates done for selection coefficient and, therefore, do not affect our conclusions in any way.

Maladapted alleles in the front end of expansion – Occasionally, simulations had the appearance of maladapted variants right on the edge of the expansion wave, as predicted by previous studies (TRAVIS *et al.* 2007; EXCOFFIER *et al.* 2009; PEISCHL *et al.* 2013). These phenomena, however, do not seem to be last very long. Especially when selection is very strong (Fig. S4), the aberrant allele frequency state does not last longer than a single generation, denoting a transient nature for the “deleterious” alleles that increase in frequency during the range expansion in our simulations.

Supplementary Tables

Table S1: Selection intensity ($\log_{10}\omega$) estimability assessment for all genetic architectures. R^2 stands for the coefficient of determination of the pseudo-observations on the parameter estimates; RMSE is relative root mean square error; and prop. HPD50% and HPD95% represent the proportion of posterior distributions encompassing the pseudo-observed value. These values were obtained based on 1000 estimates, with 1000 retained simulations out of 1 million simulations, under the stabilizing hard selection system.

Architecture	**R^2**	**RMSE***	**Prop. HPD50%**	**Prop. HDP95%**
1L2A	0.693	0.109	0.611	0.989
1L10A	0.839	0.105	0.582	0.979
10L2A	0.741	0.133	0.589	0.988
10L2A+	0.533	0.130	0.582	0.982
10L10A	0.637	0.153	0.548	0.992
10L10A+	0.839	0.113	0.552	0.982

*RMSE here is normalized by the $\log_{10}\omega$ distribution (i.e. -1 to 2)

Supplementary Figures

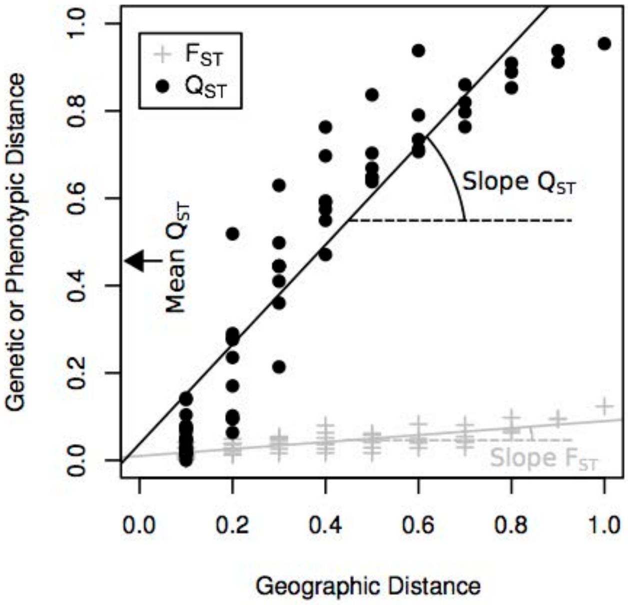

Figure S1: Pattern statistics extracted from the isolation-by-distance (IBD) pattern observed in the simulations. From the Q_{ST} IBD (black circles), we extracted **Mean Q_{ST}**, Slope of IBD (**Slope Q_{ST}**) and the sum of residuals around the linear regression (black line); for the F_{ST} IBD pattern (gray crosses), Mean F_{ST} (not shown), Slope of IBD (Slope F_{ST}) and the sum of residuals for the F_{ST} regression (gray line). Additionally, the difference between the two slopes (Δ-Slope, not shown) was also retained summing up to seven pattern statistics.

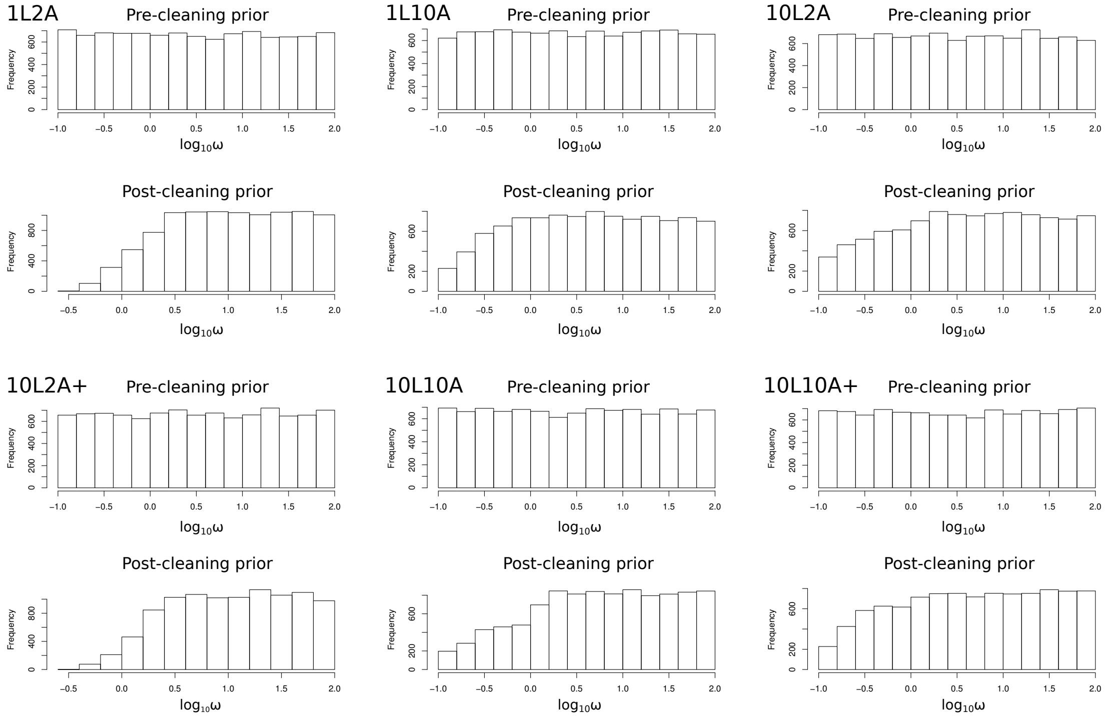

Figure S2: Comparison of pre- and post-cleaning prior distributions for selection intensity (ω) for all six architecture models in the hard selection scheme. Cleaning refers to the removal of simulations containing missing data for any of the calculated summary statistics.

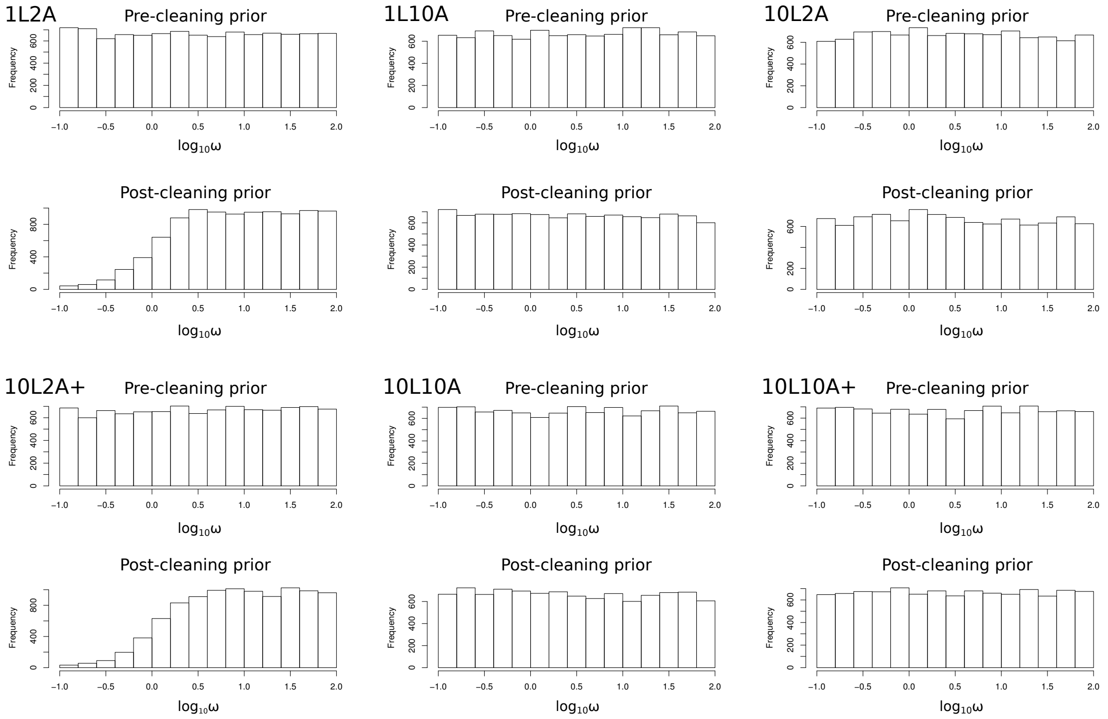

Figure S3: Comparison of pre- and post-cleaning prior distributions for selection intensity (ω) for all six architecture models in the soft selection scheme. Cleaning refers to the removal of simulations containing missing data for any of the calculated summary statistics.

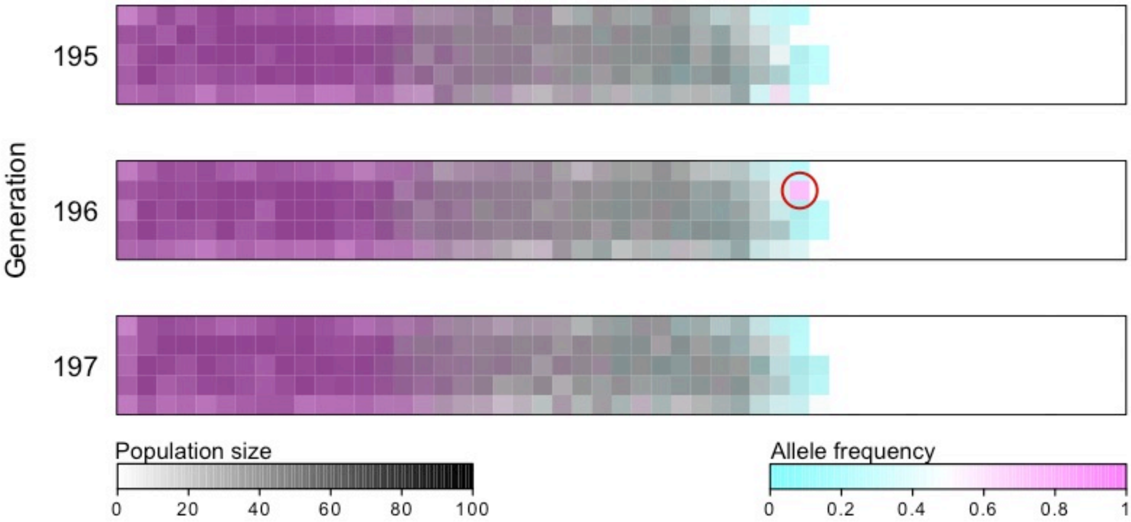

Figure S4: Deleterious mutation appearing on the front end of expansion. In a randomly chosen simulation (from generation 195 to 196), a maladapted variant appears in high frequency (magenta-colored circled deme) in a newly colonized deme. In the following generation (197), the allele frequencies of that same deme already change towards the predominance of the better-adapted allele to right-hand side of the map.

Supplementary References

EXCOFFIER, L., M. FOLL and R. J. PETIT, 2009 Genetic Consequences of Range Expansions. Annual Review of Ecology, Evolution, and Systematics **40:** 481-501.

PEISCHL, S., I. DUPANLOUP, M. KIRKPATRICK and L. EXCOFFIER, 2013 On the accumulation of deleterious mutations during range expansions. Mol Ecol **22:** 5972-5982.

TRAVIS, J. M., T. MUNKEMULLER, O. J. BURTON, A. BEST, C. DYTHAM *et al.*, 2007 Deleterious mutations can surf to high densities on the wave front of an expanding population. Mol Biol Evol **24:** 2334-2343.

General Discussion

In this thesis, all the questions investigated involved the use of computer simulations. It could not be different: All scenarios explored here involved complexities beyond any analytic approach. But not only that, the simulation approach allowed us to explore these complex models in many different ways and in greater depth. For instance, in chapter one, we used two rounds of simulations to investigate the fit of our model to the real data of human genetic diversity. In the first round, we assessed the parameter values underlying the system extracting information out of millions of simulations; and in the second, we used full genetic information out of a smaller number of simulations to try and recreate complex signatures of genetic differentiation. In the second chapter, again two rounds of simulations were used. First, we determined the neutral demographic history of the barn owls and their recolonization of Europe; second, we investigated – within the boundaries set by first set of simulations – how likely it was to neutrally evolve a color cline like the one observed in the natural populations. And it turned out that neutral evolution alone was not able to explain the observed pattern. The last chapter is a simulation-only study. There, inspired by the barn owl case above, we explored the efficiency of selection to generate an adaptive cline in range-expansion scenarios, and how could one detect and estimate selection in these cases. Curiously enough, our simulations suggest that adaptation in range expansions is detectable regardless of other parameters that could disturb this assessment. Considering that range expansions are ubiquitous in nature (Hewitt 2000; Parmesan and Yohe 2003), this approach provides a promising picture for future studies on the barn owl and other organisms, as well.

Even though the three chapters composing this thesis may look disconnected at a first glance, I believe there is an observable natural development along their execution: Going from neutrality to natural selection. In the first chapter, we explored a purely neutral demographic picture of modern human evolution. As mentioned in the chapter's text, we believe this model may be a robust alternative to other models that have been assumed as representative of the background demography in studies looking for phenotype-genotype associations and signatures of selection in humans. The second chapter showed a further development towards selection, where we devised what can be called a neutrality test for the evolution of a phenotypic trait (e.g. coat color) under a spatially and demographically explicit model of evolution. In the third chapter, we finally implemented selection to our models and verified that it can be detected and estimated, but by always keeping in mind that neutrality is the null hypothesis when it comes to investigating evolutionary questions (Kimura 1984).

However complicated the models we implemented here may seem, they were conceived to be as simple as possible. Fundamentally the only increments, when compared to previous similar studies (Fagundes et al. 2007; Antoniazza et al. 2010), are the spatially explicitness-related features. Such features go from the simulations' implementation in a two-dimensional setting to the incorporation of summary statistics that also account for spatial organization. Other studies have already applied the spatial dimension to the different related contexts (Ramachandran et al. 2005; Ray et al. 2005; Currat et al. 2010), but these were normally limited in either how geographically accurate the spatial implementation was or in the statistical treatment of the models. Here, we observed that the addition of the spatial dimension to our models, in general, brought considerable insight on the processes under investigation.

Drift: the null hypothesis

The concept of the null hypothesis is central to the scientific method. In the hypothesis formulation step, it stands for the default position, the expectation when nothing out of the ordinary is involved in the process under study (Fisher 1935). When investigating an evolutionary process, the null hypothesis is always that of neutrality (Lande 1977). But it is not always straightforward to properly address this neutral null hypothesis. The background demographic history of any system may be rather complex, leading to all sorts of observed patterns. This variety of patterns often poses a challenge to researchers looking for signatures of selection (Excoffier et al. 2009). Therefore, it is essential to any study looking for selection to implement the relevant demographic model.

Here, we addressed a particularly problematic demographic scenario in all three chapters: range expansions. This demographic process of increasing the area that a given population inhabits is intimately linked with the phenomenon of allele surfing. Surfing happens on the front end of the range expansion, where the series of founder events resulting from the persistent colonization of new locations may easily generate allele frequency clines (Klopfstein et al. 2006). To properly assess selective processes happening in such scenarios, one must explicitly take into account the demography and, since this is a spatial process, it must be modeled in a spatially explicit manner. Furthermore, as seen in chapter three, the incorporation of space into the summary statistics seems to be an important improvement as well.

In the third chapter, we make the point that selection is effective in range-expansion scenarios. This does not mean, at all, that allele surfing is in any way less likely to happen, though. Since the seminal works of Edmonds et al. (2004) and Klopfstein et al. (2006), it has been well theoretically established that the allele-surfing phenomenon is possible and very likely to happen, as supported by several other subsequent works (Hallatschek et al. 2007; Travis et al. 2007; Excoffier and Ray 2008; Hofer et al. 2009; Peischl et al. 2013). So, if there is selection in the system, it has to present a more consistent pattern than allele surfing. Using this rationale, the probability of selection can then be assessed via a simple neutrality test. In this test, the observed pattern (phenotype or genotype) is compared with a null distribution generated based on neutral simulations. Based on how far the observation is from the distribution of values generated by the neutral model, one can determine how probable it is for simple neutrality to explain the observed pattern (as done in chapter two). Allele surfing remains the most likely explanation for allele frequency clines formed in range expansions: random genetic drift (probably with some background selection) remains as the null hypothesis.

Perspectives

The most immediate and obvious future development of the line of work presented here is the application of the method of estimating selection (in chapter three) to the example of the European barn owl (chapter two). There are different ways in which the detection and (to a smaller extent) measurement of selection have been tackled in the recent literature. For example, Gutenkunst et al. (2009) present a method that uses simulations to generate allele frequency spectra, which are then compared to real data to infer past demography and selection – e.g. (Yi et al. 2010; Ellison et al. 2011). However, the models that can be implemented in this software (∂a∂i) are limited to three simultaneous populations with selection acting only at one single locus in a rather simple implementation. More complex selection regimes can be implemented in MSMS (Ewing and Hermisson 2010) that can then be coupled with an ABC analysis pipeline for the inference of the underlying parameters. MSMS does not allow the application of spatially explicit simulations, though. Simulations in space have been implemented in only a few programs. SPLATCHE (Currat et al. 2004) is arguably the most popular implementation of spatially explicit coalescent simulations, presenting a wide range of settings related to migration regimes on a two-dimension lattice. This last software, however, does not deal with any sort of selection. Yet another alternative approach towards the investigation of natural selection with simulations is the use of time-sampled data, as applied to a drug-resistance study in influenza virus (Foll et al. 2014), which takes into consideration the changes in allele frequencies across time to assess the effect of natural selection on different loci. Temporally spaced data however is not always available and, despite the recent advancements in ancient DNA analysis (Gilbert et al. 2005), will probably continue to be exception rather than the rule for natural populations. Therefore, none of the abovementioned approaches would be able to deal with the case of the European barn owl, either because of their own limitations, or because of the limitations of the data itself.

Therefore, in order to properly assess selection in case of the barn owl, one needs to implement the sort of simulations devised in the third chapter, but in a more complicated setting. If directly based on the geographically explicit simulations used chapter two, the new set of simulations will have to deal with many more demes than the ones present in the simplified implementation. Furthermore, the complex selection scheme applied to generate a gradient of selection requires the simulations to be run forward in time, discarding the choice of using faster coalescent simulations. As a result, more computational power will surely be required, leading to more processing time and larger memory consumption. Since an ABC approach often requires hundreds of thousands to millions of simulation replicates, one may have to consider some simplifications, or even a reduction of needed simulations for the estimates.

Studies in humans may also benefit from the implementation of the methods developed and used here. There have been many different approaches towards the detection of selection in human populations and different researchers are choosing a few different directions. One very popular path that has been followed is to explore the plethora of genomic that has been produced by the several population genomic projects – e.g. HGDP-CEPH (Cann et al. 2002), HapMap (The International HapMap Consortium 2003), 1000 Genomes (The 1000 Genomes Project Consortium 2010). Most of the studies looking at these databases are searching for genomic signatures of natural selection, among which the most popular is a selective sweep (Sabeti et al. 2006). Sweeps are characterized by a consistent reduction of diversity around a certain locus on the genome. This is due to the rapid fixation of one allele at that specific locus that also drags with it the linked loci that are nearby. Recombination, of course, erases this signature as one moves away from the locus under selection. This kind of signature has been detected at many locations in the human genome (The International HapMap Consortium 2005), but its accuracy in actually reflecting selective events is still debatable both because of occasional false positives (Jensen et al. 2005) and because selection does not necessarily generates a strongly marked (i.e. hard) selective sweeps (Hernandez et al. 2011).

Yet another line of evidence being used in human populations, and more related to the approaches used in this thesis work, is the use differentiation measurements (by means of F_{ST}) calculated along the genome to detect regions that are either strongly or weakly differentiated across different populations (Nielsen 2005). Even though there have been consistent advances in the analytical methods used to assess these F_{ST} signatures (Foll and Gaggiotti 2008; Narum and Hess 2011), these approaches still appear to be greatly affected by the underlying demographic history of the population being studied (Excoffier et al. 2009). And, when it comes to human populations, the matter of the fact is that they have undergone a very complex demographic history with an incredible expansion of their range since they left from eastern Africa to colonize the whole planet (Cavalli-Sforza et al. 1994; Ray et al. 2005; Fagundes et al. 2007). So, explicitly using this demographic background in future studies should provide an even more efficient way to avoid mistakes due to, for example, false positives. Besides, humans offer a huge amount of genetic data that is largely underused, leaving room for many more studies based solely on the analysis of data already produced. Simulation-based studies may profit a great deal by comparing their theoretical predictions and outcomes to the abundant observed data for human populations worldwide.

The research field of phylogeography – which concerns itself with the interplay among population genetics, phylogenetics and geography – has developed from simple mitochondrial-DNA diversity analysis considering its spatial distribution in its very origin (Avise et al. 1987), to the incorporation of formalized theoretical background, but still with a strong ad-hoc component in its analyses (Avise 1998; Edwards and Beerli 2000), to the arrival of model-based inference with hypothesis testing, in what has been dubbed *statistical phylogeography* (Knowles 2003; Beheregaray 2008; Knowles 2009; Beaumont et al. 2010). Statistical phylogeography has obvious benefits to extract from spatially explicit models. In this approach, simulations play a big role in establishing the hypotheses to be tested (normally with ABC techniques, see 'General Introduction'). Few studies, however, have done this in a spatially explicit way (Knowles 2013). Furthermore, the recent advances in landscape genetics (Manel and Holderegger 2013) and environmental niche modeling (Guisan and Thuiller 2005) provide phylogeography with even better tools to define and test hypotheses concerning past and current populations' distributions and their effects on the populations' genetic composition. The incorporation of geography – by means of spatial explicitness – seems to be the next natural step in the movement of making phylogeography a more statistically sound discipline.

Conclusion

In summary, simulations offer a wide range of possibilities in the development of many of the research areas related to population genetics. Simulations have already proven to be very useful since the late 1950’s, with the pioneering work of Alex Fraser (Fraser 1957a, b, 1958, 1959a, b, 1960) and James Stuart Barker (Barker 1958a, b), and they continue to expand in importance in the field to this date (Arenas 2012). In this thesis I hope to have accomplished two goals. First, I expect to have contributed with deeper knowledge about natural processes ongoing in humans, in the European barn owl and, potentially, other populations that have undergone range expansions. In particular, how the interplay between neutrality and selection happens in these especial demographic conditions, showing that natural selection can operate despite the complications generated by the intensified genetic drift on the edge of the expansion. Second, I hope to have produced good examples of the use of simulations – spatially explicit in particular – as tools applied to evolutionary biology. Simulations are, indeed, a powerful way to address complicated questions in science, taking advantage of the processing capacity of computers to enable scientists to look into the detailed mechanisms of nature.

General References

Aigner, M. and G. Ziegler. 2001. Buffon’s needle problem. Pp. 125-128. Proofs from THE BOOK. Springer Berlin Heidelberg.

Antoniazza, S., R. Burri, L. Fumagalli, J. Goudet, and A. Roulin. 2010. Local adaptation maintains clinal variation in melanin-based coloration of European barn owls (Tyto alba). Evolution 64:1944-1954.

Arenas, M. 2012. Simulation of molecular data under diverse evolutionary scenarios. PLoS computational biology 8:e1002495.

Avise, J. C. 1998. The history and purview of phylogeography: a personal reflection. Molecular Ecology 7:371-379.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific Phylogeography - the Mitochondrial-DNA Bridge between Population-Genetics and Systematics. Annu Rev Ecol Syst 18:489-522.

Balloux, F., H. Brunner, N. Lugon-Moulin, J. Hausser, and J. Goudet. 2000. Microsatellites can be misleading: An empirical and simulation study. Evolution 54:1414-1422.

Banks, J., J. S. Carson II, B. L. Nelson, and D. M. Nicol. 2005. Discrete Event System Simulation. Prentice Hall, Upper Saddle River, USA.

Barker, J. S. F. 1958a. Simulation of Genetic Systems by Automatic Digital Computers III. Selection between Alleles at an Autosomal Locus. Australian Journal of Biological Sciences 11:603-612.

Barker, J. S. F. 1958b. Simulation of Genetic Systems by Automatic Digital Computers IV. Selection between Alleles at a Sex-Linked Locus. Australian Journal of Biological Sciences 11:613-625.

Beaumont, M. A., R. Nielsen, C. Robert, J. Hey, O. Gaggiotti, L. Knowles, A. Estoup, M. Panchal, J. Corander, M. Hickerson, S. A. Sisson, N. J. R. Fagundes, L. Chikhi, P. Beerli, R. Vitalis, J.-M. Cornuet, J. Huelsenbeck, M. Foll, Z. Yang, F. Rousset, D. Balding, and L. Excoffier. 2010. In defence of model-based inference in phylogeography. Molecular Ecology 19:436-446.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025-2035.

Beheregaray, L. B. 2008. Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. Mol Ecol 17:3754-3774.

Bernoulli, J. 1713. Ars Conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis. Impensis Thurnisiorum Fratrum, Basel, CH.

Bertorelle, G., A. Benazzo, and S. Mona. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol 19:2609-2625.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd, and L. L. Cavalli-Sforza. 1991. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proceedings of the National Academy of Sciences of the United States of America 88:839-843.

Box, G. E. P. and N. R. Draper. 1987. Empirical Model-Building and Response Surfaces. John Wiley & Sons, Oxford, UK.

Burger, R. and R. Lande. 1994. On the Distribution of the Mean and Variance of a Quantitative Trait under Mutation-Selection-Drift Balance. Genetics 138:901-912.

Cahn, R. W. 2001. Computer Simulation. Pp. 465-488 *in* R. W. Cahn, ed. The Coming of Materials Science. Elsevier Science, Oxford, UK.

Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza. 2002. A Human Genome Diversity Cell Line Panel. Science 296:261-262.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. The History and Geography of Human Genes. Princeton University Press.

Chapleau, F., P. H. Johansen, and M. Williamson. 1988. The Distinction between Pattern and Process in Evolutionary Biology - the Use and Abuse of the Term Strategy. Oikos 53:136-138.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics 134:1289-1303.

Charlesworth, D., B. Charlesworth, and M. T. Morgan. 1995. The Pattern of Neutral Molecular Variation under the Background Selection Model. Genetics 141:1619-1632.

Currat, M., E. Poloni, and A. Sanchez-Mazas. 2010. Human genetic differentiation across the Strait of Gibraltar. BMC evolutionary biology 10:237.

Currat, M., N. Ray, and L. Excoffier. 2004. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Molecular Ecology Notes 4:139-142.

Davis, M. E. and J. S. Brinks. 1983. Selection and concurrent inbreeding in simulated beef herds. J Anim Sci 56:40-51.

Domingos, P. 1999. The Role of Occam’s Razor in Knowledge Discovery. Data Mining and Knowledge Discovery 3:409-425.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-1973.

Edmonds, C. A., A. S. Lillie, and L. L. Cavalli-Sforza. 2004. Mutations arising in the wave front of an expanding population. Proceedings of the National Academy of Sciences of the United States of America 101:975-979.

Edwards, S. V. and P. Beerli. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54:1839-1854.

Eldredge , N. and J. Cracraft. 1980. Phylogenetic patterns and the evolutionary process: method and theory in comparative biology. Columbia University Press, New York, USA.

Ellison, C. E., C. Hall, D. Kowbel, J. Welch, R. B. Brem, N. L. Glass, and J. W. Taylor. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proceedings of the National Academy of Sciences of the United States of America 108:2831-2836.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-2620.

Ewing, G. and J. Hermisson. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26:2064-2065.

Excoffier, L., T. Hofer, and M. Foll. 2009. Detecting loci under selection in a hierarchically structured population. Heredity 103:285-298.

Excoffier, L. and N. Ray. 2008. Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol Evol 23:347-351.

Fagundes, N. J. R., R. Kanitz, R. Eckert, A. C. Valls, M. R. Bogo, F. M. Salzano, D. G. Smith, W. A. Silva, Jr., M. A. Zago, A. K. Ribeiro-dos-Santos, S. E. Santos, M. L. Petzl-Erler, and S. L. Bonatto. 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. American journal of human genetics 82:583-592.

Fagundes, N. J. R., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier. 2007. Statistical evaluation of alternative models of human evolution. Proceedings of the National Academy of Sciences of the United States of America 104:17614-17619.

Felsenstein, J. 1976. The theoretical population genetics of variable selection and migration. Annu Rev Genet 10:253-280.

Fisher, R. A. 1935. The design of experiments. Oliver and Boyd, London, UK.

Foll, M. and O. Gaggiotti. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180:977-993.

Foll, M., Y. P. Poh, N. Renzette, A. Ferrer-Admetlla, C. Bank, H. Shim, A. S. Malaspinas, G. Ewing, P. Liu, D. Wegmann, D. R. Caffrey, K. B. Zeldovich, D. N. Bolon, J. P. Wang, T. F. Kowalik, C. A. Schiffer, R. W. Finberg, and J. D. Jensen. 2014. Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS Genet 10:e1004185.

Fraser, A. S. 1957a. Simulation of Genetic Systems by Automatic Digital Computers I. Introduction. Australian Journal of Biological Sciences 10:484-491.

Fraser, A. S. 1957b. Simulation of Genetic Systems in Automatic Digital Computers II. Effects of Linkage on Rates of Advance Under Selection. Australian Journal of Biological Sciences 10:492-499.

Fraser, A. S. 1958. Monte Carlo analyses of genetic models. Nature 181:208-209.

Fraser, A. S. 1959a. Simulation of Genetic Systems by Automatic Digital Computers V. Linkage, Dominance, and Epistasis. Proc. Int. Symp. Biomet. Genet.

Fraser, A. S. 1959b. Simulation of Genetic Systems by Automatic Digital Computers VI. Epistasis. Australian Journal of Biological Sciences 13:150-162.

Fraser, A. S. 1960. Simulation of Genetic Systems by Automatic Digital Computers VII. Effects of Reproductive Rate, and Intensity of Selection, on Genetic Structure. Australian Journal of Biological Sciences 13:344-350.

Garza, J. C. and E. G. Williamson. 2001. Detection of reduction in population size using data from microsatellite loci. Molecular Ecology 10:305-318.

Gilbert, M. T., H. J. Bandelt, M. Hofreiter, and I. Barnes. 2005. Assessing ancient DNA studies. Trends Ecol Evol 20:541-544.

Gill, J. L. 1964. Effects of finite size on selection advance in simulated genetic populations. Australian Journal of Biological Sciences 18:599-618.

Gillespie, D. T. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22:403-434.

Guisan, A. and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology letters 8:993-1009.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695.

Haigh, J. 2012. Probability: A Very Short Introduction. Oxford University Press, Oxford.

Hallatschek, O., P. Hersen, S. Ramanathan, and D. R. Nelson. 2007. Genetic drift at expanding frontiers promotes gene segregation. Proceedings of the National Academy of Sciences of the United States of America 104:19926-19930.

Handley, L. J., A. Manica, J. Goudet, and F. Balloux. 2007. Going the distance: human population genetics in a clinal world. Trends in genetics : TIG 23:432-439.

Hardy, O. J. and X. Vekemans. 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. Heredity 83:145-154.

Hastings, W. K. 1970. Monte-Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57:97-109.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, P. Genomes, G. Sella, and M. Przeworski. 2011. Classic selective sweeps were rare in recent human evolution. Science 331:920-924.

Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907-913.

Hewitt, G. M. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. Biol J Linn Soc 58:247-276.

Hoban, S., G. Bertorelle, and O. E. Gaggiotti. 2011. Computer simulations: tools for population and evolutionary genetics. Nat Rev Genet 13:110-122.

Hofer, T., N. Ray, D. Wegmann, and L. Excoffier. 2009. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73:95-108.

Hudson, R. R. 1991. Gene genealogies and the coalescent process. Pp. 1-44 *in* D. Futuyma, and J. Antonovics, eds. Oxfords Surveys in Evolutionary Biology. Oxford University Press, Oxford, UK.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-2314.

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro, and C. D. Bustamante. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401-1410.

Kimura, M. 1984. The neutral theory of molecular evolution. Cambridge University Press.

Kingman, J. F. C. 1982. The coalescent. Stochastic Processes and their Applications 13:235-248.

Klopfstein, S., M. Currat, and L. Excoffier. 2006. The fate of mutations surfing on the wave of a range expansion. Molecular Biology and Evolution 23:482-490.

Knowles, L. L. 2003. The burgeoning field of statistical phylogeography. Journal of Evolutionary Biology 17:1-10.

Knowles, L. L. 2009. Statistical Phylogeography. Annual Review of Ecology, Evolution, and Systematics 40:593-612.

Knowles, L. L. 2013. Testing the when, where, and how of divergence with species-specific predictions of genetic variation under alternative hypotheses. Pp. 1111. XIV Congress of the European Society for Evolutionary Biology. European Society for Evolutionary Biology, Lisbon, PT.

Lande, R. 1977. Statistical Tests for Natural-Selection on Quantitative Characters. Evolution 31:442-444.

MacKay, D. J. 1998. Introduction to Monte Carlo Methods. Pp. 175-204. Learning in graphical models. Springer Netherlands, Amsterdan, NL.

Manel, S. and R. Holderegger. 2013. Ten years of landscape genetics. Trends Ecol Evol 28:614-621.

Metropolis, N. 1987. The beginning of the Monte Carlo method. Los Alamos Science 15:125-130.

Molles, M. C. and J. F. Cahill. 1999. Ecology: Concepts and Applications. McGraw-Hill, Dubuque, IA, USA.

Narum, S. R. and J. E. Hess. 2011. Comparison of F(ST) outlier tests for SNP loci under selection. Molecular ecology resources 11 Suppl 1:184-194.

Neuenschwander, S., F. Hospital, F. Guillaume, and J. Goudet. 2008a. quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. Bioinformatics 24:1552-1553.

Neuenschwander, S., C. R. Largiader, N. Ray, M. Currat, P. Vonlanthen, and L. Excoffier. 2008b. Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. Mol Ecol 17:757-772.

Nielsen, R. 2005. Molecular signatures of natural selection. Annu Rev Genet 39:197-218.

Parmesan, C. and G. Yohe. 2003. Nature A globally coherent fingerprint of climate change impacts across natural systems. Nature 421:37-42.

Pearson, K. 1905. The problem of the random walk. Nature 72:294.

Peischl, S., I. Dupanloup, M. Kirkpatrick, and L. Excoffier. 2013. On the accumulation of deleterious mutations during range expansions. Mol Ecol 22:5972-5982.

R Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America 102:15942-15947.

Ray, N., M. Currat, P. Berthier, and L. Excoffier. 2005. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. Genome Res 15:1161-1167.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Hohna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology 61:539-542.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. 2002. Genetic structure of human populations. Science 298:2381-2385.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. Science 312:1614-1620.

Scholz, F. W. 2004. Maximum Likelihood Estimation. Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc.

Serre, D. and S. Pääbo. 2004. Evidence for Gradients of Human Genetic Diversity Within and Among Continents. Genome Research 14:1679-1685.

Spitzer, F. 1964. Principles of random walk. van Nostrand, Princeton.

Sunnaker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. 2013. Approximate Bayesian computation. PLoS computational biology 9:e1002803.

Tajima, F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123:585-595.

Templeton, A. R. 2009. Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. Mol Ecol 18:319-331.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789-796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299-1320.

Travis, J. M., T. Munkemuller, O. J. Burton, A. Best, C. Dytham, and K. Johst. 2007. Deleterious mutations can surf to high densities on the wave front of an expanding population. Mol Biol Evol 24:2334-2343.

Wakeley, J. 2008. Coalescent Theory. Roberts & Company, Greenwood Village, Colorado.

Wegmann, D., C. Leuenberger, and L. Excoffier. 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182:1207-1218.

Weir , B. S. and C. C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358-1370.

Winsberg, E. 2009. Computer Simulation and the Philosophy of Science. Philosophy Compass 4:835-845.

Winsberg, E. B. 2010. Science in the Age of Computer Simulation. The University of Chicago Press, Chicago, USA.

Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J. Wang, and J. Wang. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75-78.

Ricardo Kanitz

Born: July 20th 1983, Estrela/RS, Brazil

DEE, Biophore – UNIL
Lausanne 1015 – Switzerland
ricardo.kanitz@unil.ch
+41(0) 21 692 4243

Rue de la Colline, 10
Geneva 1205 – Switzerland
ricardo.kanitz@gmail.com
+41(0) 78 664 3431

Summary

I did my BSc and MSc studies in the city of Porto Alegre from 2002 to 2009. During this period, I worked on the genetics of the human colonization of the Americas and on the conservation genetics of one of the rarest extant mammals: *Cavia intermedia*. In 2009, I moved to the University of Geneva to work on the genetics of a repatriation program with the Española Galápagos tortoise. In 2010, I started a PhD at the University of Lausanne under the supervision of Prof. Jérôme Goudet working on modeling evolutionary processes in species such as modern humans and barn owls. **Skills:** Population genetics, phylogeography, phylogenetics, statistics, approximate Bayesian computation, Bash and R programing.

Education

2010 – 2014. **University of Lausanne (UNIL)** – PhD in Life Sciences (Ecology and Evolution). Title: The use of simulations in evolutionary population genetics: applications on humans (and owls). Supervisor: Prof. Jérôme Goudet.

2007 – 2009. **Pontifical Catholic University of Rio Grande do Sul (PUCRS)** – MSc in Zoology. Title: Evolutionary Biology of *Cavia intermedia* (Mammalia: Rodentia) - the Moleques do Sul endemic cavy. Supervisor: Prof. Sandro L. Bonatto.

2002 – 2006. **Federal University of Rio Grande do Sul (UFRGS)** – BSc in Biological Sciences.

Professional Experience

2010 – 2014. **University of Lausanne** – Research and teaching assistant in the Department of Ecology and Evolution (Prof. Jérôme Goudet).

2009 – 2010. **University of Geneva** – Researcher in the Laboratory of Artificial and Natural Evolution (Prof. Michel C. Milinkovitch).

Publications

Milinkovitch MC, **Kanitz R**, Tiedemann R, Tapia W, Llerena F, Caccone A, Gibbs JP, Powell JR (2012). Recovery of a nearly extinct Galápagos tortoise despite minimal genetic variation. **Evolutionary Applications 6:** 377-383.

Kanitz R, Trillmich F, Bonatto SL (2009). Characterization of new microsatellite loci for the South-American rodents *Cavia aperea* and *C. magna*. **Conservation Genetics Resources 1:** 47-50.

Fagundes NJR*****, **Kanitz R***, Eckert R, Valls ACS, Bogo MR, Salzano FM, Smith DG, Silva-Jr W, Zago MA, Ribeiro-Dos-Santos A, Santos SEB, Petzl-Erler ML, Bonatto SL (2008). Mitochondrial Population Genomics Supports a Single Pre-Clovis Origin with a Coastal Route for the Peopling of the Americas. **The American Journal of Human Genetics 82:** 583-592. ******These authors contributed equally to this work.*

Fagundes NJR, **Kanitz R**, Bonatto SL (2008). Reply to Ho and Endicott. **The American Journal of Human Genetics 83:** 146-147.

Fagundes NJR*****, **Kanitz R***, Bonatto SL (2008). A Reevaluation of the Native American MtDNA Genome Diversity and its Bearing on the Models of Early Colonization of Beringia. **PLoS One 3:** e3157. ******These authors contributed equally to this work.*

Awards and Scholarships

2012. Best Talk prize in the Symposium of Ecology and Evolution Doctoral Students (SEEDS).

2009. Swiss Confederation scholarship for study and research from the Federal Council for Scholarships for Foreign Students (CFBE).

2007. Brazilian Government scholarship for MSc studies from the National Council for Scientific and Technological Development (CNPq).

2005. Prize for best undergraduate work in Medical and Human Genetics at the 51st Brazilian Congress of Genetics (CBG).

2003-2006. Three times won the *Scientific Initiation Highlight* prize with two nominations to the *Young Researcher* award at the UFRGS's Scientific Initiation Symposium (SIC-UFRGS).