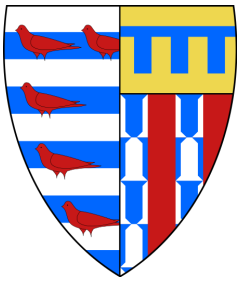
Reconstructing Population Histories in Relation to Ecology



Eleanor Frances Miller

Department of Zoology
Pembroke College

A dissertation submitted to the University of Cambridge for the degree of
Doctor of Philosophy

September 2019

To my parents, for their unwavering support

Acknowledgements

There are a huge number of people I would like to thank and without whom I wouldn't have achieved half as much as I have during my PhD.

Firstly I would like to thank both Bill Amos and Andrea Manica for all of their support and guidance throughout the last four years. Their insight, patience and enthusiasm has been invaluable.

I would also like to thank all the Postdocs and PhD Students who helped me tackle the steepest learning curve of my life. Particular thanks are due to Javier Igea who had the unenviable task of introducing me to coding for the first time, and to Pierpaolo Maisano Delser who so patiently helped me learn to navigate new languages, operating systems, programs and tools. Equally, the members of the Evolutionary Ecology Group have all been incredibly kind and supportive, always happy to respond to my silly questions, providing encouragement and support where necessary. A special acknowledgement should also go to Kathy and Corinne for the numerous tea breaks that kept me sane.

Finally, outside the department and the world of Zoology, I would like to thank the friends from home, school, and undergrad who helped me remember that the world is bigger than the Cambridge bubble. My godson Henry for bringing a new dimension into my life by being such a happy addition to the world. The Pembroke g'hals for making this Cambridge experience unbelievably entertaining. Lior, for being the best of housemates. And, last but not least, my parents, for always being there for me despite the angst! I love you. Thank you for everything.

Declaration

This thesis, entitled 'Reconstructing Population Histories in Relation to Ecology', is the result of my own work and includes nothing that is the outcome of collaboration, except that which is clearly specified in the following 'Collaborations' section. I further state that this work is not substantially the same as any that has been submitted, nor is being concurrently submitted, for a degree, diploma, or other qualification, at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit of 60,000 words for the Degree Committee in the School of Biology.

Eleanor Frances Miller
September 2019

Collaborations

Below I detail the collaborations and author contributions for each of my data chapters, without whom none of this would be possible.

Chapter 2: Published as; Miller, E.F., Manica, A. and Amos, W., 2018. Global demographic history of human populations inferred from whole mitochondrial genomes.
E.F.M. performed the analyses, prepared all figures and wrote the manuscript. A.M. and W.A. provided supervision.

Chapter 3: In prep as; Miller, E.F. and Manica, A., 2019. mtDNAcombine: tools to combine sequences from multiple studies.
E.F.M. wrote the code, performed the analyses, prepared the figures and wrote the manuscript. A.M. provided supervision.

Chapter 4: In prep as; Miller, E.F., Green, R., Balmford, A., Beyer, R., Somveille, M., Leonardi, M., Amos, W. and Manica, A., 2019. Bayesian Skyline Plots do not agree with range size changes based on Species Distribution Models for Holarctic birds
E.F.M. performed the analyses, produced the plots and wrote the manuscript. All SDMs were constructed by M.L. Expert ornithological knowledge was provided by R.E.G. who also proposed the MSI analysis. All authors provided supervision and comments on the manuscript.

Chapter 5: All SDMs were constructed by M. Leonardi. Guidance and support on bioinfomatics was provided by P. Maisano Delser. A. Manica provided supervision.

N.B. Chapters 2-4 have either been published or are in-prep for publication and only minor alterations have been made to the text of these manuscripts in order to improve readability and overall coherence of this thesis. As a result there may be some level of repetition between chapters when introducing the key topics and techniques.

Thesis Abstract

We live in an era of significant environmental and climatic change and it has even been suggested that the world is entering a new epoch, the 'Anthropocene'. To understand better how species might cope under different future climate scenarios, studies are now frequently looking to explore how they responded to rapid environmental change in the past. Whilst census data can capture contemporary trends, genetic approaches can infer population trends stretching tens, or even thousands, of years back in time.

In this thesis, I first used skyline plots to infer historical demographic trends from genetic data of a well-studied system, humans. Using this gold standard, my work revealed detailed demographic profiles, but also identified issues relating to the way key methodological assumptions are contravened. In Chapter 2 I present a discussion about the risk of misinterpretation or overinterpretation in the context of Bayesian skyline plot (BSP) analysis.

Understanding that any single profile can be problematic, when moving to non-model species, I chose to work as many species as possible. This approach exploits the recent boom in sequencing projects that has generated a huge volume of publicly available data. By building large, novel, multi-species datasets it becomes possible to construct profiles averaged over many species with similar properties, such as habitat preference. The expectation is that average profiles will prove better at capturing broad trends for the species they contain.

Collating and processing public domain data is not a trivial task. I therefore developed a pipeline, now an R package, to access and compile sequence data for over 100 species of bird, focusing on mitochondrial DNA (mtDNA). I found differences in the mean time of population expansion after the ice age between bird species associated with different habitats. However, notably, the demographic trends drawn from BSPs did not reveal a close match with the amount of available habitat indicated by species distribution models. BSPs frequently indicated population increases even though species' habitat ranges were decreasing. These results further emphasise the level of care needed when interpreting BSPs.

If genetic methods for demographic reconstruction are to be used extensively in the future, it is important that we understand what confounding factors commonly exist in real world populations so as to prevent misleading or inaccurate interpretations. To explore the impact of historic range dynamics on BSPs I created a realistic spatial demographic model for a small North American passerine, the yellow warbler (*Setophaga petechia*). From this I simulated mtDNA sequences for a number of populations across the modern species' range. With these data I'd hoped to investigate how BSP profiles varied depending on local population history. However, true demographic signals proved hard to capture and further work will be required to explore my original question more fully.

Table of contents

List of figures

List of tables

Chapter 1

General Introduction

1.1 Introduction

This is a time of great anthropogenic environmental change, the extent of which has led to the suggestion that we are entering a new human-dominated geological epoch, the Anthropocene [1, 2, 3]. It is now established that many species are declining and wide spread extinction is already a major characteristic of this era [4]. Although species extinctions have always been a feature of Earth's history, species loss is now thought to be happening ~1000 times faster than the expected 'baseline', no longer balanced by rates of speciation or repopulation [5, 6, 7]. In order to put the brakes on this ever-accelerating loss of diversity strategic conservation efforts need to be implemented, and quickly.

Predicting how species will cope in the face of a rapidly changing world, and thus how best to conserve them is, however, difficult. Whilst the population demographics of species today can be informative, this offers only a snapshot in time, and it remains hard to disentangle short-term stochastic fluctuations from longer-term trends. Yet, extensive environmental change is not novel. Over millennia, global climate oscillations have resulted in dramatic temperature fluctuations and associated large-scale changes in ice volume and glaciation. The Quaternary period, starting ~2.4 million years ago [8], saw numerous cycles of ice-ages and periods of climate warming. Initially cycling approximately every 40 thousand years, since the mid-Pleistocene, periods of glaciation have intensified, getting colder and lasting longer [9]. Now cycling roughly every 100 thousand years, these longer periods of cooling have led to the formation of larger ice sheets, reduced water availability, lowered sea levels, as well as enabling greater rates of environmental change.

In order to better understand the ways in which species are currently responding to these climatic changes, and how they might cope in the future, it is important to understand how species existed in the past [10]. In fact, the environmental oscillations of the Quaternary have provided a natural experiment with which to explore this. By mapping the demographic changes of species onto reconstructed climatic and environmental oscillations, it should be possible to build a detailed picture of past climate impacts.

1.1.1 What we know about the past

Cores of ice and sediment from around the world can be analysed for a huge range of different data that offer detailed evidence for variation in the biotic and abiotic. Climatic changes are known to be matched by changes in the chemistry of the air and water. As ice is laid down,

bubbles of air become trapped along with small particles such as ash and dust, preserving tiny quantities of the contemporary atmosphere. Once a core has been extracted, these pockets of air provide samples of historic atmosphere that can be hundreds of thousands of years old. Carbon-dating methods mean samples taken from these cores can be dated at a fine temporal resolution and so a wealth of empirical information on climatic patterns and atmospheric composition can be collected. For example, recent work by Rasmussen *et al.* [11] on three Greenland ice cores aimed to capture all the abrupt climate change events that occurred within the Last Glacial cycle at high temporal resolution. Ice cores have been drilled from ice caps and mountain glaciers across the globe but key cores include those from Greenland [12, 13], Vostok [14], and Antarctica [15, 16].

Fossilised grains of pollen found in sediment cores provide detailed insight into some of the ecological patterns and process that affected communities during glacial-interglacial stages. Plants produce pollen in large quantities and, where sediment has been continuously laid down, pollen deposits can offer an uninterrupted stratigraphic record of ecological changes. As it preserves well, in addition to insight into the qualitative changes in paleovegetation, pollen can also provide sufficient data for exploration into quantitative changes in past vegetation dynamics [17, 18]. In fact, the volume of pollen available often means a detailed, robustly dated understanding of the vegetation patterns can be obtained from a local to a continental resolution [19]. Frequently, information from sediment cores is made publicly available, adding to the ever-expanding body of data in community repositories. For example, a temporally detailed record of changes in paleovegetation throughout the Northern hemisphere during the Holocene can be drawn from data publicly accessible in the European Pollen Database (EPD) and the North American Pollen Database (NAPD) [20, 21].

A limitation of these type of empirical data is that they provide detailed description of the environment for a small number of locations, and often for a limited time period (the depth of the core). Paleoclimate models can fill these gaps by providing a coherent reconstruction of past environments (both abiotic and vegetation) through time and space. There are a number of challenges, as models are known to have biases, and often reconstructions are only available for limited time windows and geographic regions [22, 23, 24]. However, there have been a number of efforts to generate coherent reconstructions through time and space [25], and to validate them against empirical data from cores.

To gain a full understanding of the significance and impact of both historical and potential future climate change, it is necessary to be able to link the climate fluctuations to data on species' population demographics at a similar resolution. Today, by gathering empirical data on metrics such as births, deaths, and migration rates, modern ecologists can assess the health and trends in contemporary species' populations. However, insights from modern census

data are valid only for a few tens of years or, for the best-studied species, perhaps a few hundred years. These timescales are far shorter than the scales over which major climatic events occur. Therefore, most species are lacking detailed long-term population histories and, in order to explore the response of species to long term climate changes, it is necessary to use sources of data other than contemporary surveys.

Empirical evidence from sources such as the fossil record shows that the environmental fluctuations of past glacial and interglacial stages had a huge impact on population and community dynamics (e.g. [26, 27]). Large-scale climatic events offered opportunities for colonisation, adaptation, and speciation. Yet, the extent and rate of climatic change also caused extensive population displacement, declines, and extinctions [8, 28]. In northern latitudes, where substantial swathes of the area were entirely glaciated during the ice age, species were forced to undergo range contractions or shifts [29]. During interglacial phases, we know that populations recurrently colonised northern areas, tracking tolerable climatic conditions and rapidly expanding their ranges when habitat became available [28, 29]. However, to build a detailed picture of the demographic impacts for individual species through time, data with a higher temporal resolution are needed. There are several different approaches that attempt to capture the right level of resolution, inferring detailed population-size histories. These range from indirect methods that rely on proxy factors, such as changing niche space, to more quantitative methods that directly exploit genetic data.

1.1.2 Species Distribution Models to reconstruct past demographies

One powerful 'indirect' approach for reconstructing population dynamics is the use of species distribution models (SDMs). SDMs are statistical and empirical modelling tools that combine field data with environmental data to describe a species' natural distribution. Initially, SDMs were predominantly used as a descriptive tool, focused on illuminating the drivers of present-day species distributions. The development of environmental sensor technologies and methods for remote-sensing over the last few decades has, however, led to an increase in both the quality and accessibility of high resolution environmental data sets (e.g. [24]). Together with the growing number of statistical modelling methods now available, predictive statistical modelling of species' distributions is becoming more powerful [30, 31]. By assuming a known relationship between available niche space and key dynamics, such as effective population size (N_e), it is possible to reconstruct how populations might have responded to past climatic changes and predict how they may respond to events in the future.

Today, a range of different modelling algorithms are available to characterise species' natural distributions and to simulate how changes in predictor factors may have influenced available niche space through time [31]. Broadly, these modelling techniques can be cat-

egorised into two approaches, those that utilise presence-only data, and those that require presence-absence data. Presence-only data, which comprise records only of where species are found, are generally much more readily available than presence-absence data that require species being explicitly recorded as either absent or present. Reliable absence data are in fact rarely available even for well-studied species [32]. Whilst there are modelling approaches that do not require absence data at all, e.g. Environmental-Niche Factor Analysis (ENFA) [33], presence-only data techniques often use 'pseudo-absences' in their analysis. Together with presence data, these 'pseudo-absences' (points sampled at random from regions outside the area in which a species is expected to occur) can be used by established statistical tools such as generalised linear models (GLMs) and generalised additive models (GAMs) [34, 35], to predict species' distributions.

One popular approach for handling presence-only data is maximum entropy modelling (Maxent) [36] (cited $>$ 9500 times). Maxent produces a probability distribution over the study region, which is the relative probability of finding the species at any given point within the area of interest. The variance of data, such as environmental variables, from all the points where the species are recorded as present are used to constrain the predicted distribution. Any location estimated to be suitable must fit these bounds. In the final probability distribution, areas where the environmental variable values most closely match the mean values from the presence locations are given the highest probability. Despite its popularity, Maxent, like any single modelling approach, can be confounded. Ensemble methods, on the other hand, highlight areas of agreement between multiple plausible models to create a final prediction [37, 38]. This approach can therefore account for variability in different prediction methods and offer a more balanced final outcome.

No matter the model or method chosen, predictive modelling techniques come with a suite of theoretical and methodological assumptions that are rarely met with real data. Critically, it is assumed that the properties of a species-environment relationship are preserved when projecting a model through time [39, 40]. This assumption allows a species' environmental tolerance today to be used to parametrise a historical or future model. Yet, a number of factors, such as dispersal constraints, community structure, and sampling biases, can complicate niche characterisation, while biotic interactions without modern day analogues are hard to account for [37, 38, 41, 42, 43, 44]. As a result, models that attempt to estimate a species' range and abundance based solely on the reconstructed extent of available habitat may be ineffective.

1.1.3 Population genetics to reconstruct past population sizes

Another, more direct method for exploring population histories is the analysis of patterns in contemporary genetic data. At any chosen site on a species' genome, all sampled copies must be related to each other, sharing a most common recent ancestor (MCRA) somewhere in the past. This genetic relatedness between a set of samples can be used to construct a gene tree. Properties of this bifurcating tree, such as the branching pattern, will vary over time reflecting how the population size has altered. For example, looking from the past to the present, the gene tree of an expanding population will normally present faster rates of branching than a stable or shrinking population. Characteristics of genomic data can therefore act as an archive of past population dynamics that can be exploited to provide insight into historical demographic events [45, 46].

In the last few decades a swathe of statistical approaches rooted in the general principles of the coalescent theory [47, 48, 49] have been developed to explore the information captured in species genealogies (e.g. [45, 50, 51, 52]). In Kingman's coalescent [47], n samples are taken from one generation and, in the absence of selection, the ancestry of each sample is traced back by randomly selecting parents from the previous generation. When two lineages converge on the same parent a coalescent event is said to have occurred. This process of branches randomly colliding (coalescing) through the gene tree is repeated until all the branches have collapsed to one linage and the most recent common ancestor of n is found [53, 54].

In a single population, the rate at which coalescent events occur is mostly affected by the size of the population. The smaller the size of the effective population, N_e (simply put, the number of individuals in a population who contribute offspring to the next generation), the more rapidly lineages will collide and the shorter intervals between coalescences will become. At different periods, changes in population size will lead to changes in the rates of coalescence in the gene tree, from which inferences of past population dynamics can be drawn. As the coalescent only focuses on the direct ancestors of the sample and does not require all individuals in a population to be tracked back through time, this method is computationally efficient. It is important to note that there is a critical difference between the N_e of a population and the total number of individuals in a population, the census size (N). N_e is the size of a idealised population, based on simplifying Wright-Fisher assumptions (such as constant population size and random mating), which has the same amount of genetic drift as a given real population. In practise, N_e is normally smaller than N because factors such as overlapping generations, population structure, and unequal reproductive success all act to reduce the number of individuals contributing to the next generation at a single moment in time.

Originally the coalescent was used to explore species history in methods such as the variable population size coalescent model [55, 56], since then it has been extended and more complex models and methods constructed around it. As discussed by Grant [57], two key methods for historical demographic reconstructions that exploit the coalescent theory are DNA sequence mismatch analysis (MMA) [58, 59] and skyline plots [51, 60]. MMA is based on the idea that changes in population size leave recognisable signals in the distribution of pairwise differences between individuals [59]. A sudden population growth event produces a unimodal wave in the pairwise difference distribution, whilst a decline causes a 'ragged' distribution, and a stable history leaves a smoother multimodal distribution signature. As bottlenecks and the magnitude of a population expansion/contraction event also produce recognisable patterns in the distribution of nucleotide-site differences, estimates for the time of population expansion and effective population size can both be drawn using MMA [58, 61]. Now commonly implemented in the software package ARLEQUIN [62], MMA is computationally efficient but doesn't exploit all the information available in the sequence data. As a result, MMA cannot provide a temporally detailed picture of demographic changes and so it is often used in support of other methods such as skyline plot analyses.

The 'classic' skyline plot was introduced by Pybus *et al.* [50] and is based on translating the time between each coalescent event to a measure of effective population size. Having first reconstructed a gene tree from the samples, the distances between consecutive nodes in the tree, the coalescent intervals, can be defined. A piecewise reconstruction of the population's size through time can then be estimated by combining the predicted population size within each interval of the genealogy (N_i). This value is reconstructed using the relationship between the size of each coalescent interval (γ_i) and the number of lineages within that interval (i),

$$N_i = \gamma_i i(i-1)/2$$

[47–49,63]. Plotted, this profile is said to resemble a city skyline, hence the name [51]. Offering a clear, graphical way of displaying the population size information recovered from gene trees, the 'classic' skyline plot method proved popular and has prompted the development of a family of skyline methods [60, 63, 64, 65, 66]. However, like previous genealogy-based N_e reconstruction methods, the 'classic' skyline plot uses an estimated genealogy as the basis of the inference. As both topology and branch length estimates carry uncertainty this introduces error associated with phylogenetic reconstruction which is not accounted for in the output. To address this problem the Bayesian skyline plot (BSP) was developed by Drummond *et al.* [60], integrating a Bayesian framework and Markov chain Monte Carlo (MCMC) methods with the skyline concept. BSPs estimate both genealogy and population size history simultaneously and, as the posterior distribution of population size is

estimated directly from the sampled gene sequences, it is possible to include an uncertainty measure that considers confidence in both phylogenetic and coalescent values [60].

Most recently, the skyline methods family has seen the introduction of the Extended Bayesian Skyline Plot (EBSP) [66]. Where previous skyline methods were restricted to single locus data [66], the EBSP can use data from multiple loci. Compared to previous skyline plots EBSP also allows the estimated N_e to alter in a more natural, changing throughout each coalescent interval. However, this comes at a cost. As the model is not as tightly constrained, less informative datasets, such as those from single loci or with a small sample size, often mix poorly. This means analyses regularly require longer run times in order to stabilise and the rate of convergence can be low compared to BSP. Therefore, despite the introduction of the EBSP as a methodological extension, the BSP remains a popular analysis method applied to a wide variety of sequence data from many diverse taxa [67, 68, 69, 70].

Ho *et al.* [71] compare the performance of Skyline-plot methods using simulated data for two demographic scenarios, a period of exponential growth followed by a constant population size and a stable population that underwent a single instantaneous jump to a larger size. For both datasets, the BSP is able to recover the population trend, even capturing the sharp population jump in second scenario. These two modelled scenarios are, however, simplistic, involving only one change in expansion rate and / or population size. Using a combination of real and simulated single locus (mtDNA) data Grant *et al.* [72] show that, whilst BSPs are able to accurately capture changes in N_e, they are frequently blinded to events that predate any major population bottleneck. Indeed, Heled and Drummond [66] demonstrated that although BSP and EBSP agree on population trends when using the same single-locus input data, multi-locus data is key for recovering multiple population bottlenecks.

Critical to increasing the use of the skyline plots, and indeed other coalescent-based methods for demographic reconstruction, have been developments in computational power and genetic sequencing techniques. Today, there is a much wider availability of whole genome sequence data than ever before and methods have been established to exploit this resource. One such approach that has been made possible by computational and sequencing advances is sequentially Markovian coalescent (SMC) modelling [73]. Using high quality, diploid genomic sequences SMC methods can reconstruct a profile of N_e through time. The first SMC based model that explicitly aimed to explore the history of change in population size was the pairwise sequentially Markovian coalescent, or PSMC [74]. Requiring a single whole genome, PSMC uses the cross coalescent rates from two chromosomes to recover a time to the MCRA between alleles. From the distribution and variation in these timings, estimations of changes in population size can be inferred. PSMC takes advantage of both mutations and recombination events, thus maximising the amount of information available

from a genome (by contrast, skyline methods rely only on mutations, and assume fully linked markers). The multiple sequential Markovian coalescent (MSMC) [52] was later introduced, building on the PSMC framework enabling the inclusion of more than one genome. By including data from multiple individuals, more coalescent events can be captured improving the resolution. More recently the SMC++ method has been presented [75], an approach which is capable of analysing orders of magnitude more data than was possible with PSMC and MSMC.

Improving the resolution of PSMC is important if studies are interested in capturing details of population history at recent timescales (the details depend on generation times, and mutation and recombination rates; for humans, PSMC does not provide clear signals for periods more recent than 20,000-30,000 thousand years ago). At these more recent times, PSMC has reduced power because reconstructions are being inferred from the very few events that have had time to occur. Therefore, estimates of population size become unreliable with large variance. Using data from multiple individuals, as per MSMC, does allow better estimation of population demographics at more recent times [52]. However, it should be noted that scaling up to population level reconstructions from a sample size of ~1-4 individuals creates innate uncertainty as one genome only represents a single example of a population's history [76]. A notable disadvantage of MSMC is that it requires multiple high-quality phased genomes from the species of interest, data that are likely to be too difficult or costly to obtain for the majority of non-model species. Although, with the introduction of SMC++ there is no longer a requirement for phased data, there is still a requirement for a number of high-quality genomes. This volume of quality genomes does not apply to many non-model systems as yet.

Another possible approach to explore demographic histories from genetic sequence data is to use an Approximate Bayesian Computation (ABC) method. In simple terms, ABC is a generalised, simulation-based, statistical framework. ABC algorithms iterate through a series of contending hypotheses extracting a set of summary statistics from each model. The fit of different hypotheses is then compared to the empirical data with the aim of identifying which parameter configuration best describes the observed data.

Conceptually, ABC was first considered in the 1980s by Rubin [77]. The idea was further developed by Tavaré *et al.* [78] in the 1990s and, in 1999, one of the first true ABC algorithms to be extended to the field of population genetics was implemented by Pritchard *et al.* [79, 80]. Since then there have been several methodological extensions and developments of ABC methods, designed to better answer specific population genetic questions (reviewed by [80]). However, ABC remains a generic model fitting approach. Whilst it offers a very flexible framework for model fitting and has great utility for detecting a number of changes

in population size, the reliance of ABC on a few summary statistics makes it unsuitable for continuous reconstructions in the style of skyline methods.

Aside from the influence that type and quality of sequence data has on the accuracy of any coalescent-based method to accurately estimate a population history, it is critical to consider the ecological context from which the data were collected. As coalescent methods are based on evaluating patterns in gene trees, analyses can be confounded alternative factors that create similar patterns. For instance, data from structured populations must be carefully handled in order to avoid spurious results or misinterpretation of outputs. The underlying genealogy of a structured population can be similar to that of a bottlenecked or expanding population, depending on when the structuring occurred and changes in migration [76, 81]. For example, in any coalescent based analysis, a comparatively recent introduction of structure to a population can cause a false signal of population decrease [82]. However, ancient structure in the depths of a population history can create the impression of a much larger overall N_e than actually existed by acting to preserve many more deep branches in the species' gene tree. The deeper the population structure, the greater the over estimation of the effective population size will be. To minimise the impact of structure careful consideration must be given to sampling strategy and the way data from different regions or groups are integrated.

1.1.4 Different markers

Genetic techniques for demographic reconstruction all require some form of genetic marker on which to work. Fundamentally, a genetic marker simply captures patterns of accumulated differences in DNA between individuals. Over time, the increase in quality and accessibility of sequencing methods has led to an array of different choices becoming available. Some of the first markers available were protein isozymes. Isozymes work on the basis that different alleles for a functional protein will have slightly different chemical structure and so different electrophoretic mobility. Studies using isozymes were very important, helping to advance our understanding of gene flow, population structure and diversity.

As methods for DNA sequencing developed further so other classes of markers were introduced. Methods based on restriction enzyme digests, such as Amplified Fragment Length Polymorphism (AFLP) and Restriction Fragment Length Polymorphism (RFLP), were common before generally being succeeded by the use of microsatellites markers. Then, as sequencing continued to become more affordable, studies turned to directly sequencing segments of interest, targeting specific fragments or regions such as the mitochondria (mtDNA). Today, the explosion in whole genome sequencing (WGS) means hundreds of studies are now able to work with WGS methods.

Every marker has its own good and bad points and, whilst no single marker is ideal [82], arguably one marker that offers a good balance between levels of variability and ease of use is mtDNA. The high mutation rate of mtDNA [83] means that differences are accumulated faster than in other, more slowly evolving markers. An mtDNA gene tree will, therefore, yield more branches over a shorter time span and so can offer better resolution of recent population histories. Alongside this rapid mutation rate mtDNA also has a high copy number and a highly conserved gene content between species, both of which help to make it a relatively easy and cost-effective to amplify [84, 85]. The putative lack of recombination in mtDNA also means that past coalescent events can be identified without having to consider the additional complexities of genomic rearrangements [86]. This combination of utility and affordability has led to mtDNA becoming one of the most widely used markers of the last two decades.

Despite the many advantages of mtDNA, it has to be acknowledged that there are some fundamental issues with it as a marker. For instance, the improved resolution offered by mtDNA at more recent time scales is traded-off with the risk of saturation, where bases could be switching so frequently they revert to an 'original' state after a previous mutation event [87]. The primary impact of saturation would be to blind mtDNA gene trees to events in deep history meaning much older events may not reliably be recovered. Equally, the lack of recombination in the mitochondrial genome means that all sites share a common genealogy and have to be considered as one locus, yet, any single locus considered on its own might be misleading. Incomplete lineage sorting (ILS), where the history of a single gene tree differs from the overall species tree [88], is one issue that could cause a spurious history to be recovered. Equally single locus data sets are, generally, more vulnerable to loss of deep signal as they can only represent a single version of the population history. Events such as bottlenecks can cause the loss of lineages that could have provided insights into a species' evolutionary history. Major population expansions or contractions can dominate the recoverable signal in a single locus, leaving only the history of the population since that event.

As additional, independent, loci can provide more data with which to characterise an event, the use of WGS data can improve statistical power and so help to tackle some of the limitations associated with single locus studies. However, whilst the publication and availability of WGS is expanding, the volume of high-quality WGS published for non-model species remains relatively small. At the start of this PhD a number of low-quality genome sequences had been published (e.g. [89]) and since then a range of projects have started systematically to build WGS databases for different groups (e.g. [90], http://b10k.genomics.cn). Yet, traditional, single-locus, mtDNA studies remain regularly used because they still offer an affordable

and accessible approach. Indeed, Garrick *et al.* [91] demonstrate that, although recent years have seen a decrease in the proportional use of mtDNA compared to other sequence data for animal studies, the position of mtDNA as a well-established marker means mitochondrial sequence data are likely to remain a valuable part of genetic studies in the future.

Whatever the marker used, all methods of inference based on the coalescent offer temporally scaled profiles that require an accurate mutation rate in order to link it to true time. However, calculating accurate estimates of mutation rate has proven a recurrent challenge in genomics. Changes in rate are found between taxa, between regions of the genome, between loci and even between bases. It is also known that mutation rate estimates drawn from fossil records diverge from pedigree estimates [92, 93] with fossils yielding slower rates than more pedigrees. Ho *et al.* [94] suggested this relationship between the short-term mutation rate and long term substitution-rate could be captured in an exponential rate decay curve. However, if an inappropriate mutation rate is used, any reconstructed history will be discfonnected from reconstructions of other variables, such as climate. This disconnect could lead to a loss of information about demographic drivers or even cause misinterpretation of any population patterns recovered.

Every marker and method has pros and cons, and no single option offers an optimal balance for investigating all population genetics questions. However, for exploring the period of time covered by the last glacial cycle (from the start of Marine Isotope Stage 3 (MIS3) ~60 thousand years ago (kya)) a good compromise seems to be offered by using a combination of mtDNA and BSPs. Aside from the physical properties of mtDNA that make it a useful marker (e.g. fast mutation rate, high copy number, etc), a large volume of publicly available mtDNA sequences exist for an extensive range of species. This level of data availability offers an exciting opportunity to undertake large-scale comparative studies that may be able to offer a more balanced picture than any single species or single family study could achieve.

1.1.5 Birds as an indicator species

To explore the Anthropocene impact on species more broadly it is necessary to exploit information from indicator species. Indicator species are species whose population health can be used as a proxy for the condition of the habitat they live in and the health of species they live with. Birds are a diverse taxonomic group, found across a huge range of structurally diverse habitats, and are frequently key species in ecological networks. This group also tend to hold positions high in food webs and are known to be sensitive to environmental change [95]. This combination of features means that birds are often both convenient and responsive bio-indicators [96, 97, 98]. As a result, avian datasets have provided the backbone of much work on the ecosystem impacts of anthropogenic and climate change. For example, Klos *et*

al. [99] used data on bird phenology to help understand localised impacts of climate change while O'Connell *et al.* [100] used bird community composition to rank habitat condition from good to poor across the central Appalachians, discriminating between levels of human environmental disturbance.

Having been the subject of large-scale monitoring over several decades there are historical data sets available for many bird species. Indeed, demographic changes in bird populations are frequently being assessed using data from long running multinational monitoring schemes such as the Pan-European Common Bird Monitoring Scheme (PECBMS) (e.g.[101]). These schemes are also used as the basis for broader environmental and habitat health indictor systems. Herrando *et al.* [102], for instance, used existing bird monitoring data for the Mediterranean to develop an indictor methodology designed to track the impacts of land use changes on broad biodiversity trends in the region. Also, Stephens *et al.* [103] used data from the PECBMS and the North American Breeding Bird Survey (BBS) to assess the impacts of contemporary climate on the abundance of populations. An earlier version of their methodology was even included to assess progress towards achieving the United Nations Convention on Biological Diversity's Aichi targets on biodiversity. However, even data provided by long running monitoring schemes will be constrained to a narrow window in time, unable to offer any insight into the history of demographic change pre-records.

With reductions in sequencing cost and continual refinement of sequencing processes, the last few decades have seen a boom in genetic studies from non-model organisms. Indeed, aside from census, phenology, fecundity and other demographic data from surveys, avian studies have frequently collected genetic samples for different purposes. As the volume and species diversity of publicly available genetic data increases, so it becomes more and more realistic to compile large, multispecies datasets to address novel questions. In fact, studies based on genetic datasets drawn from large scale community databases are now beginning to emerge [104, 105]. In the coming years the utility of pipelines that allow exploration of inherently messy data, originally gathered in smaller subsets, will begin to grow.

1.2 This thesis

In this thesis, I explore the ability and reliability of Bayesian skyline plots for recovering population histories. Whilst several studies have investigated the performance and properties this approach using idealised simulated data (e.g. [71]), in non-model systems we frequently only have small quantities of mtDNA and the sample sizes are often not ideal. Therefore, I begin in Chapter 2 by investigating how BSPs perform on a gold standard dataset from humans, with high quality complete mtDNA sequences and a number of populations that

have been intensively sampled. Human past demography is highly studied both by geneticist and archaeologists, providing a clear context to interpret the BSPs recovered in different populations.

In Chapter 3, I then progress to develop a pipeline for downloading, collating and analysing diverse, multi-species, mtDNA data from Holarctic bird species from GenBank. Whilst this database offers a potential source of large amount of data, the quality of both individual datasets and their depositions is highly variable, and I provide a framework and related tools to synthesise this information into a coherent dataset. I use this dataset in Chapter 4 to investigate the relationship between changes in N_e as recovered from BSPs and the range dynamics that can be reconstructed from SDMs. Finally, in Chapter 5 I use a spatially explicit model fitted to genomic data from the yellow warbler, an American passerine, to explore the ability of BSPs to recover complex but realistic expansion dynamics. I hope that the work in this PhD shows the advantages, as well as some of the complexities, that come with handling large comparative datasets as well as exploring the influences and confounding factors that commonly exist in real-world data.

Chapter 2

Global Demographic History of Human Populations Inferred from Whole Mitochondrial Genomes

A version of this chapter has been published as Miller, E.F., Manica, A. and Amos, W., 2018. Global demographic history of human populations inferred from whole mitochondrial genomes. *Royal Society Open Science*, 5(8), p.180543.

Abstract

The Neolithic transition has led to marked increases in census population sizes across the world, as recorded by a rich archaeological record. However, previous attempts to detect such changes using genetic markers, especially mitochondrial DNA (mtDNA), have mostly been unsuccessful. I use complete mtDNA genomes from over 1,700 individuals, from the 1000 Genomes Project Phase 3, to explore changes in populations sizes in five populations for each of 4 major geographic regions, using a sophisticated coalescent-based Bayesian method (Extended Bayesian Skyline Plots) and mutation rates calibrated with ancient DNA. Despite the power and sophistication of this analysis, I fail to find size changes that correspond to the Neolithic transitions of the studies populations. However, I do detect a number of size changes, which tend to be replicated in most populations within each region. These changes are mostly much older than the Neolithic transition and could reflect either population expansion or changes in population structure. Given the amount of migration and population mixing that occurred after these ancient signals were generated, I caution that modern populations will often carry ghost signals of demographic events that occurred far away from their current location.

Keywords: demographic history, coalescent, mitochondrial DNA, Bayesian skyline plots

2.1 Introduction

The Neolithic transition was associated with major cultural and societal changes, and a number of archaeological lines of evidence point to a rapid increase in census population size following the advent of food production and the associated sedentism [reviewed in [106]]. However, past attempts to detect such size changes using genetic markers have generally failed to find any signal attributable to the Neolithic transition [107, 108, 109, 110]. When population changes were detected, these were generally dated to older times, leading to the suggestion that populations that later adopted agriculture might have started growing before the advent of food production [109].

A major difficulty in interpreting these results is that genetic dating of events is a very challenging endeavour, as mutation rates (which provide the molecular clock used to convert genetic changes into calendar years) come with high levels of uncertainty [111]. Over the last couple of years, the availability of ancient DNA, coupled with sophisticated tip based calibration methods that use the age of ancient samples to estimate the rate at which differences between sequences accumulate, has greatly improved the accuracy of mutation rates, especially for mtDNA [112].

Here, I take advantage of the Phase 3 data of the 1000 Genomes Project, which now includes over 2500 individuals from several major continental regions [113]. I use Extended Bayesian Skyline Plots (EBSPs) in BEAST to best reconstruct the changes in effective population size through time, and take advantage of leaf-calibrated mutation rates based on extensive data from ancient DNA [112, 114, 71, 115]. While this work builds on several previous analyses that are conceptually similar e.g. [107, 108, 116], the current study includes a number of important technical advances that should improve the ability to detect any demographic signal of the Neolithic transition that might be present. Furthermore, compared to previous analyses based on the Phase 1 1000 Genomes data, it is now possible to include five South Asian populations, sequenced as part of Phase 3.

2.2 Materials and Methods

2.2.1 Sampled populations

The Phase 3 sequence data from 20 populations, comprising 5 populations for each of the 4 main geographic regions of Europe, East Asia, South Asia and Africa, were downloaded from the 1000 Genomes Project website [www.1000genomes.org/data,[113]], including whole

mitochondrial genome data for 1,999 individuals. I decided not to analyse populations from the Americas due to the region's complex history of admixture [117, 118].

The European populations were; Finnish sampled in Finland (FIN); European Caucasians resident in Utah, USA (CEU); British in England and Scotland (GBR); an Iberian population from Spain (IBS) and Toscani from Italy (TSI). Representing East Asia were the Han Chinese in Beijing (CHB); Southern Han Chinese (CHS); Dai Chinese from Xishuangbanna, China (CDX); Kinh population from Ho Chi Minh City, Vietnam (KHV) and Japanese from Tokyo (JPT). The South Asian populations were Punjabi Indians from Lahore, Pakistan (PJL); Gujarati Indians in Houston, USA (GIH) as well as Indian Telugu sampled in the UK (ITU); Bengahli from Bangladesh (BEB) and Sri Lankan Tamil from the UK (STU). Finally, in Africa, I chose a population from the Western Division within The Gambia (GWD); Mende from Sierra Leone (MSL); the Yoruba from Nigeria (YRI); the Esan, also from Nigeria (ESN); as well as the Luhya from Webuye in Kenya (LWK). Full details of the populations and the original sampling and sequencing methods can be found on the 1000 Genomes Project website (www.1000genomes.org).

2.2.2 Data partitioning

Mutation rates of mtDNA vary among bases according to region, codon position and depending on whether the region is genic or non-genic [119]. The power of this analysis was maximised by accounting for these heterogeneities using the partitioning scheme developed by Rieux *et al.* [112], who used PartitionFinder [120] on a large panel of modern and ancient complete mtDNA genomes. Following their best model [112], the partitions, substitution model and rates were: the hypervariable segments 1 and 2 (HVS1+HVS2) with a TN93+I+G substitution model and a rate of 31.434 x10^{-8} μ/Site/Year; rRNA and tRNA (r+tRNA) with TN93+I+G and 1.007 x10^{-8} μ/Site/Year; protein coding positions at 1st and 2nd codon (PC1+PC2) with TN93+I+G and 0.756 x10^{-8} μ/Site/Year; and protein coding positions at the 3rd codon (PC3) with TN93+G and 3.323 x10^{-8} μ/Site/Year. See Supplementary Fig. A.1.

2.2.3 Data analysis

I analysed the mtDNA data with the Extended Bayesian Skyline Plot (EBSP) method, a Bayesian, non-parametric technique for inferring past population size fluctuations from genetic data. Building on the previous Bayesian Skyline Plot (BSP) approach, EBSP utilises a piecewise-linear model and Markov Chain Monte Carlo (MCMC) methods to reconstruct a populations' demographic history [66] and is implemented in the software package BEAST v2.3.2 [115]. Alignments for each of the 20 populations were loaded separately into the

Bayesian Evolutionary Analysis Utility tool (BEAUti v2.3.2) in NEXUS format. BEAUti is a graphical user interface that supports the creation of BEAST XML input files, enabling the user to easily set parameters and specific model criteria. Within BEAUti, a *'Gamma Category Count'* of four was selected for partitions using +G models to allow for the inclusion of gamma rate heterogeneity. For partitions using +I models, the *'Proportion Invariant'* was set to 0.1 and the 'estimate' box selected allowing the analysis to include a proportion of invariant sites. *'Coalescent Extended Bayesian Skyline'* process was used and the *'Population Model'* population factor set as 0.5 to account for the female only contribution to the N_e [66]. A linked, strict, molecular clock and linked phylogenetic tree was used for all analyses. All other operator settings were left as default.

Each population was run separately with each run consisting of 100 million generations sampled every 10,000 steps and the first 10 million samples discarded as burn-in [66]. To maximise comparability, the sample size used was 85 for all populations, equal to the smallest sample (MSL). Where more samples were available, a random 85 were selected. Each data set was subject to two replicate runs to confirm repeatability. Runs were analysed using Tracer v1.6 and convergence was verified by plotting MCMC chain traces and ensuring that the effective sample sizes (ESS) of all relevant parameters exceeded 200. Independent runs were then combined using LogCombiner (v2.3.2) and again analysed using Tracer (v1.6) to determine that the same stationary distribution was sampled both times. Demographic reconstructions were then plotted in R (v3.2.3).

To confirm that 85 samples provide adequate data for accurate population reconstruction, I re-ran the analyses using all available samples for the population from each major region with the maximum samples. Run length was extended to 200 million generations to account for increased sample size, while burn-in remained at 10%. For the four major regions these largest samples were: IBS in Europe IBS ($n = 107$ samples); CHS in East Asia ($n = 105$ samples); GIH in South Asia ($n = 103$ samples); GWD in Africa ($n = 113$ samples). The resulting profiles were essentially identical, though with somewhat narrower confidence intervals (Supplementary Fig. A.1). Consequently, for maximum comparability, the results presented are for the sample size of 85 that could be achieved for all populations.

Each BEAST analysis yields a profile comprising 85 paired size – time estimates that together describe the demographic history of that population. Unfortunately, different populations have different history lengths and the densities of the points vary along each profile. To attempt to obtain a fair estimate of similarity between any given pair of profiles, I used the following strategy. Comparisons were made based on 20 evenly-spaced time intervals summing to the length of the shorter history (i.e. $0, L/20, 2L/20, \ldots L$, where L is maximum age-point of the population with the shorter history). At each of these 21

time-points, the size of each population was estimated using linear interpolation between the two immediately flanking values, and the total difference calculated as

$$\lambda = \sum_{i=0}^{i=20} |\log(s1) - \log(s2)|$$

where s1 and s2 are the interpolated sizes in the two populations at bin i.

2.3 Results

The effective sample size (ESS) of relevant parameters was greater than 200, my criterion for convergence, for 19 populations. One South Asian population (BEB) failed to reach 200, so the results for this population should be treated with some caution. However, since replicate subsets all yield similar profiles and the average profile is similar to others from the same geographic region, I believe that the broadly correct demographic history has been recovered. A constant population size can be confidently rejected for all 20 profiles as the 95% highest posterior density for the number of population changes excludes 0 in every instance.

In terms of population similarity, I used autosomal SNP data from the 1000 Genomes Phase 3 to calculate Fst between all population pairs, using the method of Hudson *et al.* [121]. As expected, the major geographic regions are clearly resolved (Supplementary Fig. A.2). In addition, I also compared the similarity of the demographic profiles obtained using BEAST. Here again, populations from the major geographic regions tend to form discrete clusters (Supplementary Fig. A.3). Such clustering is consistent with the idea that populations from the same part of the world tend to have experienced similar influences on when and how much they increased in size.

2.3.1 Regional demographic histories

Africa: Profiles for the five African populations are presented in Figure 2.1. As with all other regions, graphs are arranged to correspond approximately to their geographic locations. All African populations share a large, stable ancestral size that shows little change in the East (Luhya) and an expansion in the West. The signal of expansion is stronger and starts later (around 10-11 kya) in the Nigerian populations Esan and Yoruba compared to the Mende and Gambian populations whose expansion initiates closer to 18 kya. As such, the four West African populations, particularly the Nigerians, echo the profiles found in Southern European populations (see below), albeit with a significantly larger initial size.

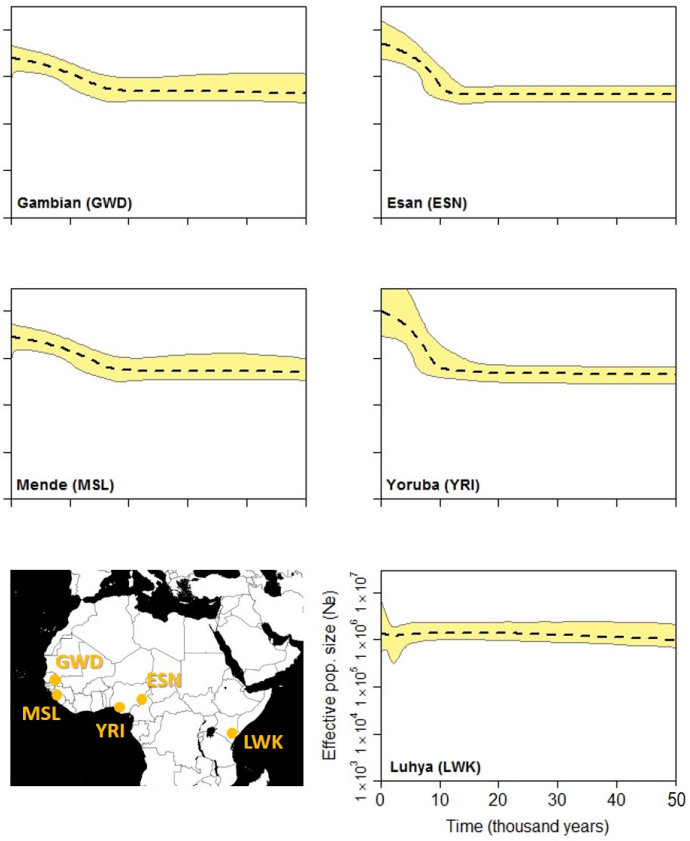


Fig. 2.1 Extended Bayesian Skyline Plots (EBSPs) for five Africa populations. Each separate population history is inferred from 85 full mitochondrial genomes. Dotted line is the median estimate of effective population size (N_e) and the thin grey lines show the boundary of the 95% central posterior density (CPD) intervals. The x-axis represents time from the present in thousands of years. All plots are on the same scale. Map labelled with geographic origins of sampled populations.

Europe: The five European profiles are presented in Figure 2.2. The four southerly populations all show profiles with a stable size up to ~14 kya followed by a sudden, rapid increase that becomes progressively less steep towards the present. There is also a north-south trend, with confidence intervals becoming broader towards the north, particularly for the oldest time-points. The Finnish population profile appears rather different but this is to be expected both because it is so far north and because previous studies have identified Finns as a strong genetic outlier in Europe [122, 123, 124, 125].

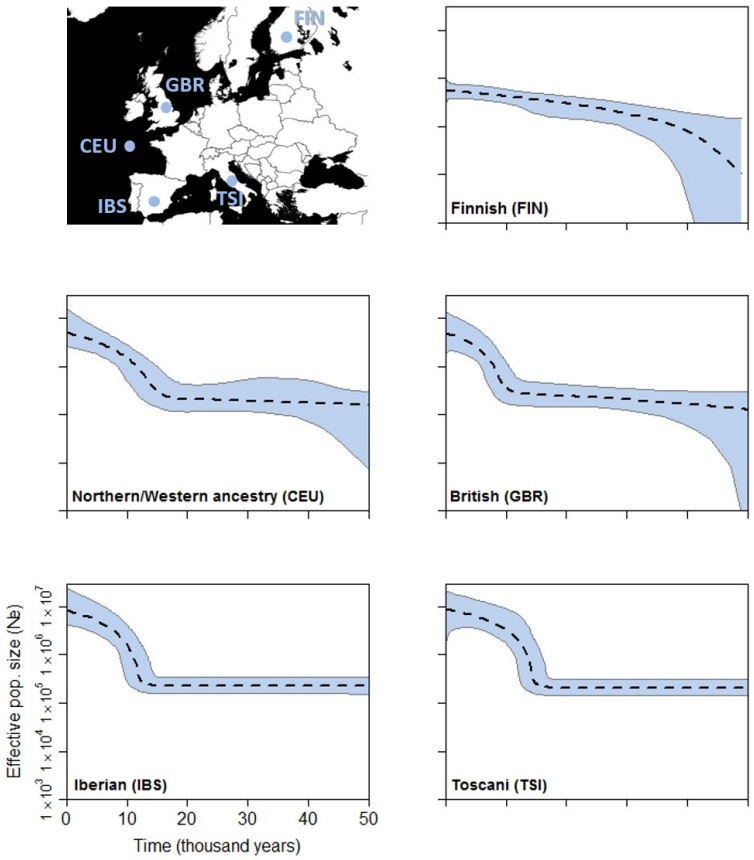


Fig. 2.2 Inferred demographic histories of five European populations. Dotted line is the median estimate of N_e and the thin grey lines show the boundary of the 95% CPD interval. The x-axis represents time from the present in years and all plots are on the same scale. Map shows origins of sampled populations.

South Asia: The five profiles for South Asia are shown in Figure 2.3. All populations reveal a period of rapid growth ~45 – 40 kya which then slows. Near the present the two southerly populations, GIH and STU both show evidence of a decline. However, this may be due to these samples being drawn from populations no longer living on the sub-continent, with the downward trend capturing a bottleneck associated with moving to Europe / America, perhaps accentuated by the tendency for immigrant populations to group by region, religion and race [126].

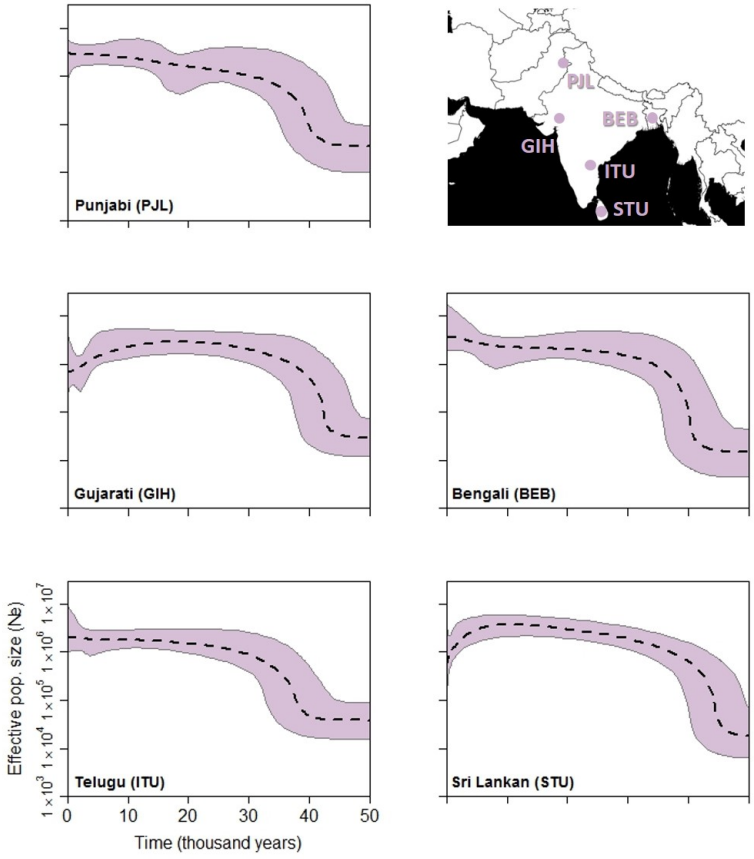


Fig. 2.3 Inferred South Asian population demographic histories. Dotted line is the median N_e estimate and the thin grey lines show the boundary of the 95% CPD intervals. The x-axis represents time from the present in thousands of years and all plots are on the same scale. The map shows location of sampled populations.

East Asia: The five population profiles for East Asia are presented in Figure 2.4. All five profiles show a generally upward trend with variable confidence limits suggesting unresolved demographic complexity. The two south-eastern populations, Dai and Kinh, share similarities with the South Asia group, having a rather rapid increase around 45 kya. The other three populations show a weaker initial expansion but instead show some similarity to the European populations in terms of a recent accelerated expansion before or around 10 kya. This secondary expansion appears to begin a little later in Japan, as observed by Zheng *et al.*. [107].

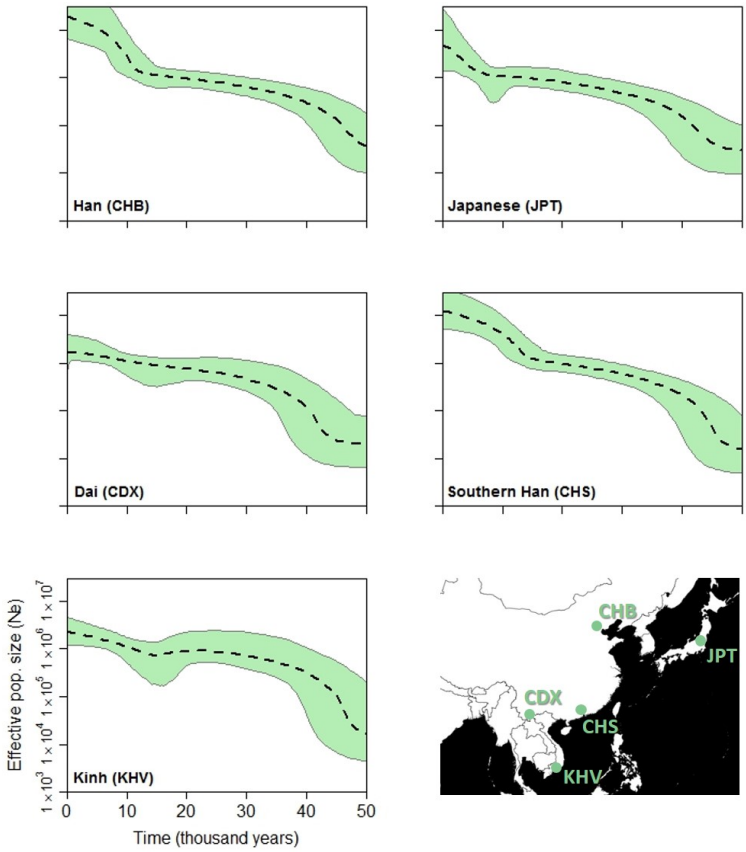


Fig. 2.4 Individual EBSPs of the five East Asian populations. Dotted line is the median estimate of N_e) and the thin grey lines show the boundary of the 95% CPD intervals. The x-axis represents time from the present in thousands of years. All plots are on the same scale. Map shows populations sampled.

For a more objective depiction of the extent to which profiles are more similar between related populations I plotted a measure of curve similarity (CS) against Fst (Figure 2.5). Since CS captures differences in both size and profile shape, it is not surprising that the values found are highly variable. Nonetheless, curve similarity does increase with Fst and CS values tend to be more similar to each other with particular region-region comparisons

compared with the overall range. Thus, South Asian profiles seem to have relatively less affinity to Europe and Africa yet greater affinity to East Asia.

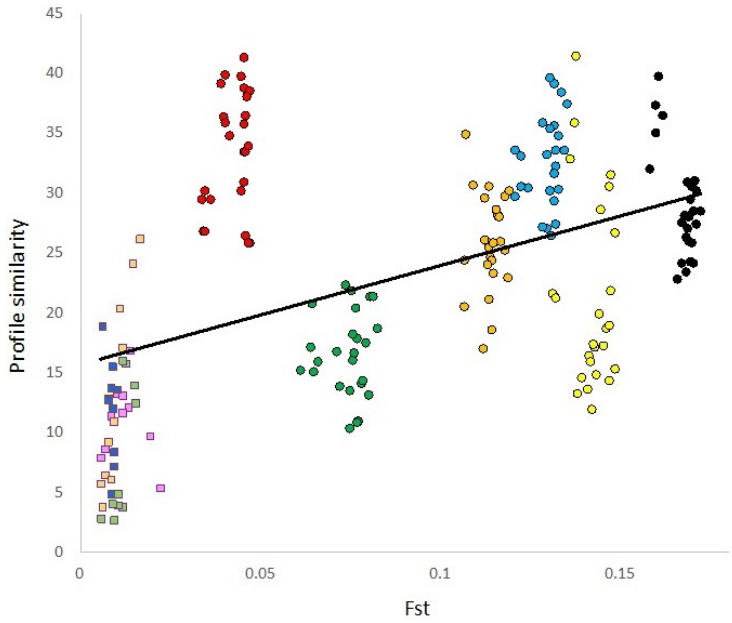


Fig. 2.5 Relationship between profile similarity and genetic distance, measured as Fst. Comparisons between regions, circles, are colour-coded: black = AFR-EA; yellow = AFR-EUR; blue = AFR-SA; orange = EUR-EA; green = EA-SA; red = EUR-SA. Comparisons within regions, squares, are coded: peach = EUR; pink = EA; dark blue = EA; light blue = AFR. Profile similarity is calculated as l summed over 20 evenly spaced intervals (see 2.2 Materials and Methods).

2.4 Discussion

I used the Bayesian program BEAST to infer population histories for 20 global human populations using whole mitochondrial genome sequence data from Phase 3 of the 1000 Genomes Project. This analysis builds on earlier studies using the Phase 1 data e.g. [107, 108] or single haplogroups e.g. [125, 127]. The Phase 1 data lack any South Asian populations and include several American samples with complex patterns of European admixture [117, 118]. By moving to Phase 3 data I have been able to increase greatly the number of within region comparisons. I show that populations from the same region show greater similarity between their demographic profiles than populations from different regions. There is also a tendency within each region for the profiles to exhibit geographic trends. Whilst I was able to detect changes in population sizes in all 20 populations, all these increases appear to be

too old to represent the effect of the Neolithic transition, in line with previous analyses of more limited datasets [107, 108, 109, 110, 127]. See Supplementary Fig. A.2.

Compared with previous studies, this analysis has been able to deploy larger sample sizes, a coalescent model and improved mutation rate estimates. The fact that I still fail to detect a clear signal from the Neolithic transition, may suggest that even complete mtDNA genomes lack sufficient resolution to detect changes over this time scale. This conclusion agrees with simulations by Aimee and Austerlitz [128], who argued that only microsatellites, which evolve appreciably faster, might offer sufficient genetic resolution to detect such a recent event. Having said this, there may be factors other than sheer mutation rate that confound the ability to detect recent trends. For example, the mitochondrial genome is only a single marker and hence, by chance, may fail to capture signals seen in gene trees produced from other markers. Thus, Silva *et al.* [127] analysed population samples from South Asia, combining autosomal and Y-chromosome markers to reveal patterns consistent with sex biased dispersal. Here, the lack of a signal of population expansion in mtDNA reflected demographic changes associated with males rather than insufficient mitochondrial mutations.

The possibility that different markers can tell different stories is emphasised by the work of Karmin *et al.* [129]. Their results for mtDNA data are broadly similar to mine, with populations in Africa showing gradual increase over time, an early expansion in Asia and more recent expansion in Europe. However, they use Y-chromosome markers to detect a population reduction in the mid-Holocene, a trend that I fail to detect. One possibility is that the prevailing population structure resulted in relatively stable female effective population size at a time when sex-specific drivers acted to reduce the male N_e.

Verifying the ability of programs like BEAST to infer accurate population histories by simulation is difficult. Modern human populations have extremely complicated histories with changing levels of substructure, stratification by religion and politics and mixing through trade, wars and slavery [130, 131, 132]. Yet, at the same time, some level of constancy is maintained through the persistence of insular minority groups. Such complexity seems too great to be captured convincingly by simulations. Consequently, one of the best ways to show success of the method is through the consistency of profiles obtained from independent samples collected from related but distinct populations. The fact that I find profiles that are more similar to each other within a region but differ between regions therefore gives me confidence that I am are picking up genuine regional differences: populations that are nearer geographically are more similar in terms of their inferred demographic history, captured more objectively in the general positive trend between Fst and profile similarity. In turn, this pattern also indicates that a sample size of 85 individuals is adequate data for

accurate population reconstruction, something I further confirmed by extensive re-running with different, randomly selected subsets.

The reconstructed population profiles generated here exhibit several features that appear consistent with known demographic events. Thus, the very early expansion observed in East and West Asian populations is compatible with the out of Africa bottleneck and subsequent expansion. Similarly, the timing of the expansion in Southern Europe could be seen as pointing to the beginning of the Neolithic Transition in the Near East, the source of farmers who later colonised the rest of Europe [133, 134, 135]. Alternatively, it is possible that the European expansion dates reflects a population expansion from source regions into the area as the ice retreated at the end of the last glaciation, In either case, the expansion signals reflect older events that likely happened before the lineages arrived at where they were sampled. Equally, the profiles and expansion dates found across South Asia are similar to those recovered in previous studies [127, 136] such as work by Silva *et al.*, who suggest that the expansion signal seen in their BSPs around 45-35 kya may be indicative of a secondary founder event in the region that obliterated more ancient signals.

Within each major region the profiles are generally rather similar, though interestingly there also appear to be east-west / north-south trends. Thus, in Africa, the two westernmost populations GWD and MSL both show an earlier but smaller expansion compared with the two Nigerian populations YRI and ESN. Similarly, among the East Asian populations there is tendency for the most recent expansion to occur more recently in the more northern populations CDX, CHB and JPT. It is also noticeable that the more northern / eastern South Asian populations have profiles that are most similar to the more western East Asian populations, with PJL and ITU appearing most similar to CDX and KHV. These putative trends require a further increase in sample size to quantify but suggest that demographic change can in principle be tracked across both time and space.

The fact that the earliest signals are found in populations that are mostly far from where they were when the changes occurred raises an important cautionary note in interpreting these trajectories: such reconstructions are only valid under the assumption of a closed population [71, 137, 82]. Population structure, expansions and mixing all generate apparent changes in N_e which might have nothing to do with actual changes in the local census population. A single uniparental marker offers a powerful tool for investigating demographic histories but interpretation must be done carefully with the understanding of what details might be missing, wiped out or swamped by a suite of different influential processes [127]. This issue is not specific to BEAST, and other approaches such as PSMC suffer of the same limitations [76]. It is this need to avoid likely admixed populations that caused me to exclude populations from the Americas.

When interpreting any demographic reconstruction based on the coalescent it is also important to remember that accuracy of any dating will depend on the quality of the calibrating information provided, in a BSP analysis this calibration information is provided by the mutation rate. Whilst I used the best available mutation rate [112], calculation of these values is complex and uncertainty remains high. Altering the mutation rates used would re-scale the plots and may, in turn, alter the interpretation. For example, faster rates of mutation than were used in this study might have brought the timings of the expansion events found in line with the dates of the Neolithic transition, whilst slower rates might have pushed the expansion events further back in time, in line with an expansion after the end of the Last Glacial Maximum (LGM), ~21 kya.

In conclusion, expansion of the analysis of the 1000 Genomes Project mitochondrial DNA data to Phase 3 allows novel comparisons both between and within four major geographic regions. Although it remains difficult to ground-truth the dates, the fact that clear geographic trends are apparent suggests that the relative size and timing of expansions found are likely reliable. However, this shows that, even with the gold standard of data, BSPs can be misleading. Naïve interpretation of the data would imply that the populations studied all experienced expansions that initiated prior to the adoption of agriculture. It was previously suggested that such changes might be associated with changes in lifestyle, such as an increase in sedentism, that occurred before the advent of food production in the Neolithic [106]. Rather, I suggest that the signal from each local population in fact reflects a much deeper demographic history, not from those local derived populations being studied, but older 'source' populations which underwent geographic expansions. To what extent complexity of interpretation is due to humans having a very extreme expansion dynamic or not, it is hard to tell. Exploring this in more detail will be the aim of the rest of my thesis.

Chapter 3

How to Build a Comparative Dataset from Existing Sequences

Abstract

Today an unprecedented amount of genetic sequence data is stored in publicly available repositories. For decades now, mitochondrial DNA (mtDNA) has been the workhorse of genetic studies, and as a result, there is a large volume of mtDNA data available in these repositories for a wide range of species. Indeed, whilst whole genome sequencing is an exciting prospect for the future, for most non-model organisms' classical markers such as mtDNA remain widely used. By compiling existing data from multiple original studies, it is possible to build powerful new datasets capable of exploring many questions in ecology, evolution and conservation biology. One key question that these data can help inform is what happened in a species' demographic past. However, compiling data in this manner is not trivial, there are many complexities associated with data extraction, data quality and data handling. Here I present the mtDNAcomp package, a collection of tools developed to manage some of the major decisions associated with handling multi-study sequence data with a particular focus on preparing mtDNA data for Bayesian Skyline Plot demographic reconstructions.

Keywords: demographic history, R, mitochondrial DNA, Bayesian skyline plots

3.1 Introduction

Understanding a species' demographic past can help inform many questions in ecology, evolution and conservation biology. Consequently, there is a lot of interest in methods that are able to infer how a population's size may have changed through time. Traditional methods relied on insight from the fossil record [138, 26, 139]. However, although fossils are informative about many species, including our own, they remain a limited resource with coarse geographic and temporal resolution. In contrast, genetic methods have the potential to offer better resolution and are now established as the primary means by which a population's distant past can be interrogated.

Mitochondrial DNA (mtDNA) has been used widely for demographic reconstruction. The haploid nature of mtDNA along with its rapid rate evolution [83], lack of recombination [140] and uniparental mode of inheritance [141] make it more sensitive to capture changes in population size than slower evolving nuclear genes [142] (Supplementary Fig. B.1). MtDNA therefore has the temporal resolution to capture the impacts of relatively recent events that might be of interest, such as the Last Glacial Maximum (LGM). In combination with coalescent-based reconstruction methods such as Bayesian Skyline Plots (BSPs) [60], mtDNA can be used to estimate a detailed population profile that stretches back tens, or even hundreds, of thousands of years. On the negative side, since the mtDNA genome does not recombine, it acts as a single locus and thus is subject to high levels of stochasticity, necessitating larger sample sizes of individuals than if multi-locus data were available.

With the falling costs of whole genome sequencing (WGS) and the growing interest in large scale sequencing projects, such as the Bird 10,000 Genomes Project (B10K) [90], the availability of WGS data is rapidly increasing. Using a single, high quality, diploid genome sequence, the pairwise sequentially Markovian coalescent (PSMC) method [74] can reconstruct a profile of population size through time for that species. However, PSMC is limited in its ability to capture details of population history more recently than ~1,000 generations ago [52]. The multiple sequential Markovian coalescent (MSMC), a method that builds on the PSMC framework, somewhat resolves this issue, using data from multiple individuals to improve the resolution of PSMC by an order of magnitude to more recent times [52]. However, this method is costly, requiring multiple, phased, high-quality genomes from the species of interest. Whilst phasing data may get easier as average sequenced read lengths increase, this is still a non-trivial step and phased data is frequently too difficult or costly to obtain for non-model species.

Whilst WGS is an exciting prospect for the future, for most non-model organisms' classical markers such as mtDNA remain widely used [91]. Indeed, the falling costs of high throughput DNA sequencing, coupled with routine deposition of project data into public

databases such as the National Centre for Biotechnology Information's (NCBI) GenBank [143], has created a burgeoning resource of mtDNA sequence data. For the first time, these databases contain sufficient sequence data to allow users to build quality meta-datasets. Although individual studies may only be able to undertake spatially and temporally restricted sampling efforts, by creatively using pre-existing resources from multiple studies, it is now feasible to improve sampling strategy, range coverage and sample sizes without additional sampling. As the workhorse of population genetics studies for many decades, public domain mtDNA data are available in large numbers for a wide range of species across most higher taxa.

Although sequence databases are normally curated, data input is generally not standardised or error checked. Studies differ greatly in the length and identity of target sequence, the quality of sequence curation and, while some studies upload all sequences obtained, others merely upload unique haplotypes. There are also instances of incorrect sample assignation. Altogether, this means that to compile a comparable set of sequences from multiple studies requires extensive data processing. In this chapter, I consider the practicalities and problems faced by a meta-analysis of publicly available data and present the mtDNAcomp package. The mtDNAcomp package is a collection of tools developed to manage some of the major decisions associated with handling multi-study sequence data with a particular focus on preparing mtDNA data for BSP population demographic reconstructions (Figure 3.1.).

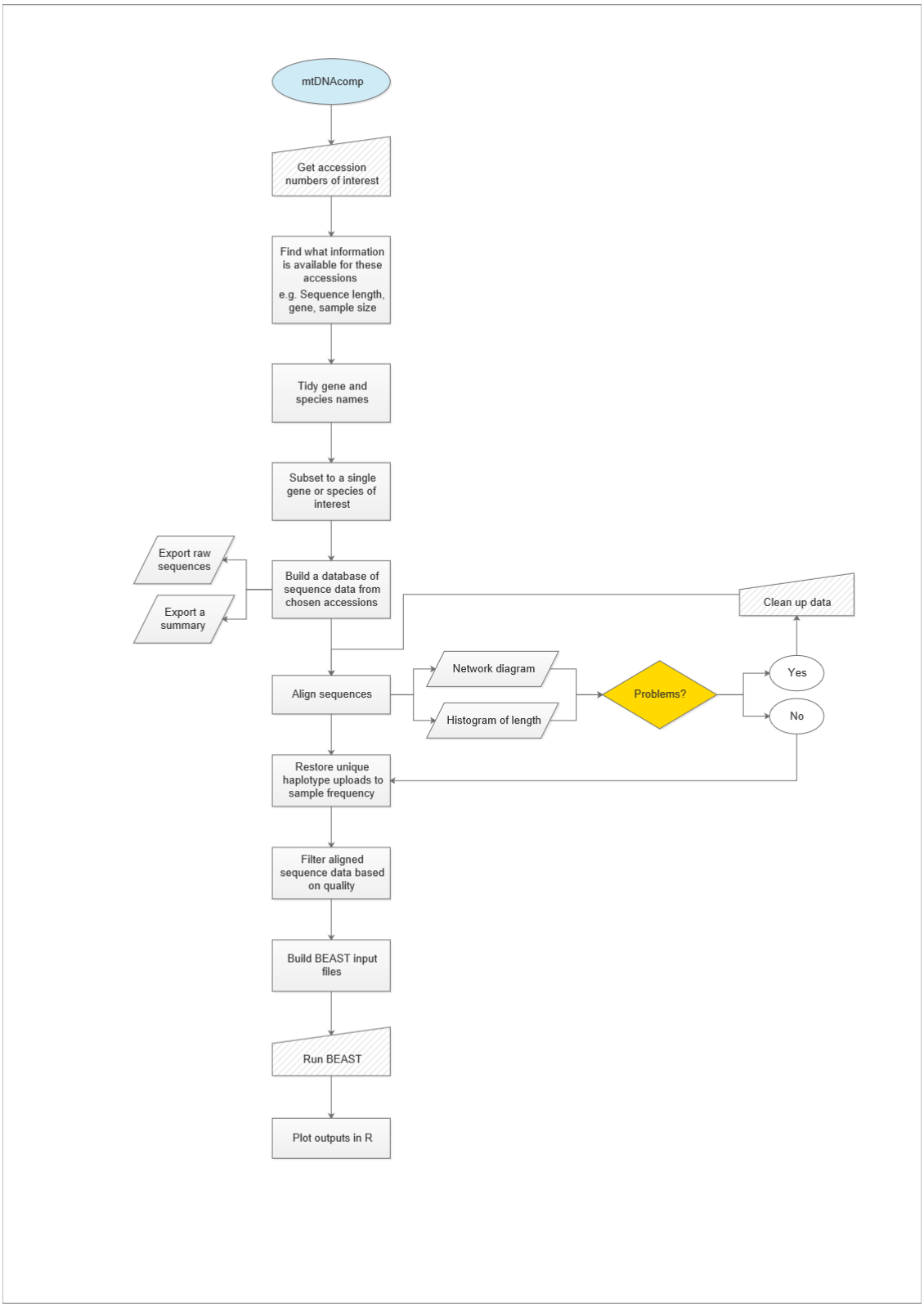


Fig. 3.1 Flow diagram of mtDNAcomp pipeline showing decisions and steps supported by the package.

3.2 Materials and Methods

3.2.1 Data preparation

Raw data

Step one is to search annotated DNA databases to determine how many data sets are available. I focus on GenBank, which is the main public repository for mtDNA datasets. Their website is intuitive, and it is easy to set up a search for a given taxon. In mtDNAcomp, information is imported (e.g. title of associated paper and sequence length) about relevant accessions into a dataframe with the 'build_GB_dataframe' function. I then proceed to explore and clean up this information to make it comparable across studies, and thus allow data for the any given species to be merged and comparable datasets for multiple species created.

It should be noted that, although GenBank staff review all submissions to GenBank, and quality control checks are performed before release, there is no standardised format for entering descriptive information. As a result, features such as alternative abbreviations for gene names, deprecated species names, subspecies names, and simple misspellings are all common. When nomenclature does not match between entries, filtering a large database for comparable samples is complex so, the mtDNAcomp pipeline includes two functions ('standardise_gene_name', 'standardise_spp_name') that allow the user to re-set common alternatives / errors in species and gene names to a chosen standard value.

Avoiding duplicate sequence entries

As BSP analysis draws information from haplotype frequency, it is important to try to avoid inclusion of duplicate entries because these can skew estimates of effective population size (N_e) and alter the reconstructed timings of demographic events. Repeated entries for a single sample can come from multiple sources, for example, the NCBI Reference Sequence (RefSeq) project [144] aims to curate records and associated data, providing a set of reference standards. As these data are drawn from the International Nucleotide Sequence Database Collaboration (INSDC, which consists of GenBank, the European Nucleotide Archive (ENA), and the DNA Data Bank of Japan (DDBJ)) databases, a basic search can recover two accessions for the same sample; the RefSeq accession and the source record(s). In this instance, the duplicates can be distinguished because all RefSeq records begin "NC_", while simple repository accessions never include an underscore. The code ('load_accessions' function, called within 'build_GB_dataframe') will automatically (and silently) remove any RefSeq record if the original accession is also found to be present in the dataset; however, users should be aware that these exclusions are being made.

Duplications can also arise from re-uploaded / re-sequenced samples. This occurs most frequently when multiple studies sample a single museum specimen, though there are other scenarios which can lead to a single individual being sequenced by multiple studies. Re-sequenced samples are often hard to identify and recognising repeated use of published alternative ID numbers (such as specimen numbers) are sometimes the only indications that the same individual has been sequenced by multiple studies. Although an occasional duplicate entry in a moderate sample size of around 100 sequences is unlikely to cause a significant skew in the recovered population history, authors should be conscious that this source of duplicate entry exists and needs to be avoided whenever possible. Unfortunately, there is no simple programmatic way to avoid it given the information provided in GenBank.

Alignment

After sequence data have been obtained, they must be aligned. A number of public domain software programs are available that can achieve this, including T-Coffee [145], MUSCLE [146] and MAFFT [147]. In mtDNAcomp, I chose to use ClustalW [148], implemented through the R package *msa* [149]. Though BEAST can handle missing / ambiguous bases [150], I consider it best to use alignments without gaps or ambiguities. Whilst some insertions or deletions may be genuine, when working with sequences from multiple sources, the data are likely to have been sequenced with different techniques to varying standards. Inclusion of basic sequencing errors could drive miscalculations in later analyses and the volume or type of errors will not be consistent across all studies, nor across all taxa. It is therefore recommended that, to ensure consistent sequence quality, all sites with ambiguities, insertions, deletions and missing data should be removed. This is done automatically within the 'align_and_summarise' function in mtDNAcomp.

Diagnostic plots

Compiling data from multiple studies produces a series of known challenges which will be tackled individually in the following sections. The 'align_and_summarise' function draws a series of key diagnostic plots for each species dataset being handled. These plots are designed to help the user quickly visualise the data, enabling rapid identification of any problems in the aligned data. If these diagnostic plots look problematic, it is then possible to return to the original input files and revaluate the raw sequence data on a case-by-case basis. The user can then decide to proceed with the analysis, return to the pipeline with an edited set of samples, or choose to drop the dataset entirely if too many samples / studies have to be excluded.

Sequence length

For any group of studies there will be numerous reasons the samples were original collected and sequenced. Each project will have had, among other things, a different budget, time constraints, target area of the mitochondrial genome, and available sequencing technology, meaning that different lengths of the genome / target gene will have been sequenced. In some instances, only very short sections of the gene of interest will have been sequenced. If the number of base pairs (bp) is too low, the sample is unlikely to hold enough information to be informative for population demographic reconstruction. The 'align_and_summarise' function will drop individual accessions that are below a user-set threshold before processing the data. There can be no out-of-the-box value for this 'minimal length' as the most appropriate size will vary with a wide range of factors such as the gene under investigation, mutation rate, absolute gene length, and the available sample size. However, excluding any samples that clearly hold insufficient information before aligning and cropping sequences to the maximum overlapping area prevents an excessive loss of information if one very short sequence were included.

Equally, above the minimal length that has been set, there can still be a wide variance in the number of base pairs, or region of the focal gene sequenced by different studies. Automatically cropping all the sequences to the maximum overlap length may result in the loss of a large amount of data unbeknownst to the user. Therefore, in order that the process of alignment and sequence trimming is transparent, one of the diagnostic plots mtDNAcomp produces is a histogram showing the original variation in sequence length as well as the length of the trimmed, maximum overlap, dataset (Figure 3.2, Appendix B, vignette section 'Diagnostic plots'). This plot flags instances where a large number of base pairs have been removed in order to include a shorter sequence. Sequence length versus sample size is a trade-off that individual users may want to weight differently depending on the data available. By presenting the information, mtDNAcomp allows the user to go back, review, and revise the input data if they want.

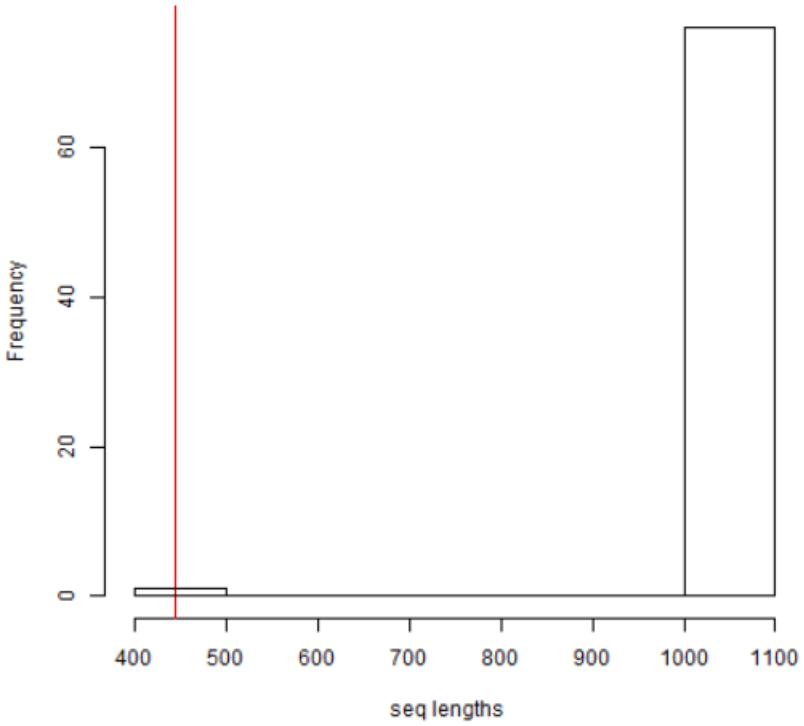


Fig. 3.2 Example of diagnostic plot for sequence trimming in the 'align_and_summarise' function. Histogram shows that, in order to trim all sequences to the maximum overlapping length (red line), the majority of samples have had to be heavily cropped.

Haplotype frequency

Studies differ in the ways they deposit data. Some upload a single copy of each haplotype they found, while others upload sequences for each individual sampled. Datasets built exclusively of unique haplotypes are not suitable for a BSP analysis [57]. Where only unique haplotypes have been uploaded, it is vital to find the number of samples these haplotypes represent, or the study must be excluded. Routinely checking every source publication to see whether they uploaded only a single copy would be tedious and may become impractical for larger analyses. To guide this process, the 'align_and_summarise' function flags studies in which all haplotypes are unique (i.e. there are no replicates) as candidates for further investigation. A text file of individual accession numbers is also produced, including a column for the user to input new frequency information. Once satisfied that the sampled frequency for each haplotype has been recorded correctly within this document, the table can be read back into R, and the function 'magnify_to_sampled_freq' will build the dataset up to correct sample sizes. See Appendix B, vignette section 'Haplotype frequency', for a worked example.

Population structure

Population sub-structure is known to cause problems for demographic reconstructions methods and BSP analysis is no exception [82, 151, 152]. BSP analysis, like other coalescent methods, is founded on the Wright-Fisher model and hence assumes panmixia [51]. This assumption is violated by population sub-structure [82, 153], which acts to reduce the probability that lineages from different demes coalesce. In practice, depending on the sampling strategy employed, sub-structure can lead to inflated population size estimate in older parts of the reconstructed history but can also noticeably reduce apparent population size at the present [82]. Accurate demographic reconstruction therefore requires careful consideration of whether sub-structure is or might be present.

Once DNA sequences have been identified, downloaded, aligned, and multiplied up to sampled frequency, the level of population structure can be assessed. One of the most intuitive approaches is to visualise the haplotype network diagram for each dataset. To maintain a streamlined approach, network diagrams are drawn within R using the package *'pegas'* [154]. These network diagrams are one of the diagnostic plots created by the 'align_and_summarise' function (Appendix B, vignette section 'Network diagram').

Depending on the level of supplementary detail available for each sample, the decision to split a population for analysis can be simple. For example, in instances where sampling location data are available and clear geographic divisions coincide with major genetic clades, datasets can be separated and multiple sequence files handled as individual datasets.

However, it is important not to over-split the data. Clades are a natural feature even of fully homogeneous populations, so if any obvious clades are removed, what is left will tend to be star-like haplotype clusters. Such clusters will often yield a signal of population expansion which may or may not be real. Deciding if and where to divide datasets remains one of the more subjective and difficult challenges and it can be worth investing time into running data sub-sets to determine the impact of alternative splitting decisions.

Outliers

I frequently found instances of extreme outliers, single haplotypes that were separated from all others by many base changes. Such outliers may be genuine but equally may reflect immigrant individuals, sample mislabelling [155], amplification of integrated nuclear copies, incorrect accession codes, or even result from poor-quality sequencing. The benefits of including these outliers in case they are genuine are far outweighed by the risk that they distort the process of inference. I therefore recommend that outliers are identified and removed, although it is useful to retain copies of the original files so that the impact on inferred demographic histories can later be investigated if necessary. Within the 'outliers_dropped' function, any "extreme outliers" are removed from the working dataset. I recognise that factors such as species life history, species population history, data availability, and data quality will influence the criteria for data inclusion. Therefore, the degree of separation from other haplotypes necessary for a sample to be classified as an "extreme outlier" is something that can be set by the user.

3.2.2 Setting up and running BEAST

BEAST input

In large comparative studies, as many steps as possible should be kept constant. This minimises the chance that the analysis becomes prohibitively time-consuming and helps to make the outputs as directly comparable as possible. The process of setting up and parameterising a BSP analysis in BEAST is well-described in several papers as well as in the accompanying textbook [150] so I will not go into detail here. Briefly, BEAST requires values for a range of parameters of which arguably the most important is mutation rate. Selection of an appropriate mutation rate is a persistent problem in genetic studies. With BSP analyses, mutation rate influences the scaling of both inferred population size and timing of events, but it does not affect the overall profile shape. Both the mutation rate itself and its associated confidence will vary between taxa and it is necessary for the user to consider how best to standardise this to maximise consistency across profiles. For certain groups,

attempts have been made to provide rates for a large number of taxa [156], though this kind of resource is far from universal as yet.

To maximise the probability that a given run converges, it can be a good idea to use fairly tight constraints on initialising parameters such as the number of population size changes. This decision will be study-specific with no one-size-fits-all approach. Moreover, changing priors and parameter values can alter outputs and should be done in accordance with best Bayesian practices [150]. Bearing this in mind, I suggest that a loss of resolution in some profiles may be a necessary trade-off if the maximum number of species is to be included.

The mtDNAcomp package function 'setup_basic_xml' utilises the *'babette'* package [157] to build basic XML files from the data processed earlier in the pipeline. The skeleton XML files will need editing (e.g. defining mutation rate, model choice, output names) but their creation minimises the number of steps the user needs to perform manually, speeding up the process and reducing the opportunity for the introduction of human error. Once parameterisation decisions have been made and the XML input files finalised, whenever possible, I encourage use of the *BEAGLE* library [158] when running BEAST2, since this can significantly improve the speed of a run.

BEAST output

Interpretation of BEAST outputs has been covered well in the literature e.g. [72, 82] and by those who designed and built the software [116, 114, 66, 159, 115]. As with any statistical model, checks need to be done to confirm the reliability of the output. In BEAST2 these are generally undertaken using the software package Tracer [160] and focus on appropriate convergence of the Markov chain. As a rule of thumb, outputs should be treated with caution wherever the effective sample sizes (ESS) for a given parameter drops below 200. Similarly, duplicate runs should be used to confirm that the posterior probability distributions stabilise at similar values. Whilst ESS values can be captured directly through the package *'babette'* [157], I believe that a visual inspection of each run in Tracer is best practice. Whilst doing so, it is then possible to export extensive summary data from the 'Bayesian Skyline Reconstruction' tab (found under 'Analysis' in Tracer). These Tracer exports are detailed, informative, and concise to work from, ideal for tasks such as downstream data visualisation as is done in mtDNAcomp.

Plotting profiles in R

BSPs can be drawn using the programme Tracer [160]. However, for more flexibility, and to facilitate exploration of the profiles in greater detail, I chose to visualise the reconstructed

profiles in R. Within the mtDNAcomp package vignette (Appendix B), I present example code for plotting Tracer output data as BSP profiles (section 'Exploring outputs'). However, it is anticipated that data presentation will be highly project specific, therefore this code is not tied up in functions, enabling easy editing and adaptation by the user.

Cautions

Skyline plots offer a powerful tool set but are easily over-interpreted. Although covered in several recent reviews [72, 57], over-interpretation continues to be an issue and hence its dangers are worth re-iterating. Unsurprisingly, problems are greatest with weaker data: smaller sample sizes, uneven sampling strategy, and / or when drawn from a species with strong population substructure [57, 82]. For example, an investigation of the same species, the common rosefinch, based on two mtDNA datasets with very different sample sizes gives us contrasting results (Figure 3.3). The smaller sample set, cytb, suggest a weak linear increase in size over time but the larger dataset, ND2, uncover a rapid, almost 100-fold increase in size. This clearly indicates that interpretation of BSP plots must be done with appropriate consideration for the data quality.

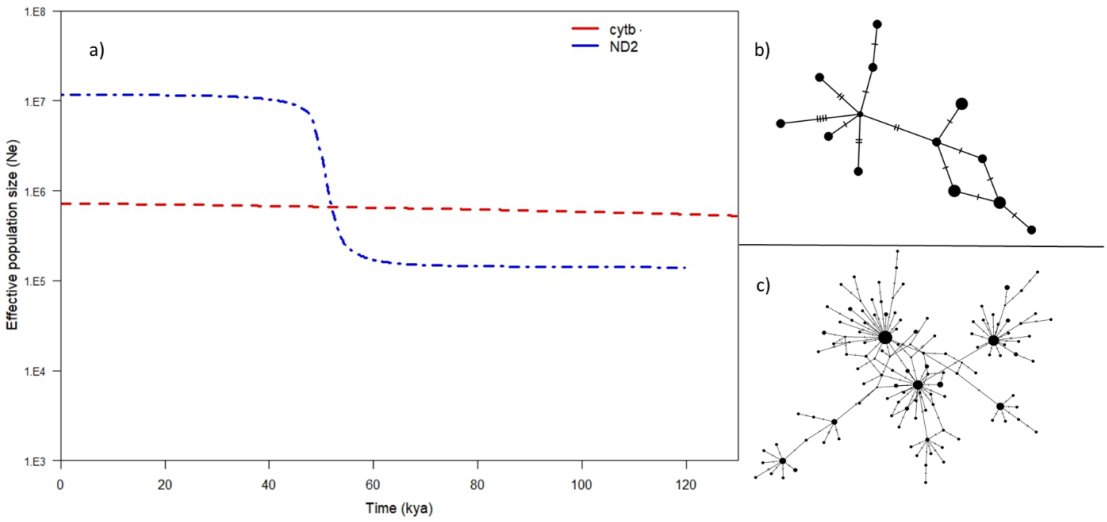


Fig. 3.3 Comparison of two dissimilar BSP profiles drawn from different mtDNA datasets of the common rosefinch. a) Red line is median value for cytb BSP profile, blue line is median value for ND2 BSP profile. The cytb dataset includes 15 samples, ND2 dataset 190 samples. The varying levels of information available for inferences to be drawn from are clearly shown in b) the median joining network (MJN) for cytb dataset, and c) MJN for ND2 dataset.

Uploading sequence data

When assembling large annotated DNA databases using published data, many sequences are 'lost' due to inaccuracies or inconsistencies in how the data are uploaded to repositories. Unless the accession process becomes more standardised, idiosyncrasies and errors will continue to render an appreciable proportion of the potential data unusable. I therefore encourage people who wish to upload data to take the time to complete as many supplementary fields as possible and to be sure they undertake basic formatting checks such spell-checks, correct capitalisation and use of standard abbreviations. Where accompanying information is not uploaded to repositories, I urge authors to make this information easily accessible to readers. For example, downstream use will be facilitated by providing haplotype frequency data or detailed sampling location data as supplementary files (ideally well formatted text files which are easy to process) rather than embedded tables or images within manuscripts.

3.3 Conclusions

With the exponentially expanding volume of data in public DNA sequence repositories, there is now more genetic information available than ever before. Building large meta-data sets by combining existing data offers the opportunity to explore new and exciting avenues of research e.g. [161, 162, 163]. However, compiling multi-study datasets still remains a technically challenging prospect. Unknown sequence quality, little to no control over sampling structure, potential errors in species identification, and limited control of sample size are all factors that can negatively affect a comparative study if not carefully handled.

Here I present the mtDNAcomp package, providing a pipeline to streamline the process of downloading, curating and analysing mitochondrial sequence data (Figure 3.1). At the moment, the lack of standardisation in the data upload process exacerbates the inevitable complexities of combining data from multiple origins. Whilst some samples, sequenced early in the molecular era, are allowably poorly documented I urge people to be careful when uploading data today. The more information about a sample that is included online, alongside sequence data, the more likely that sequence will be usable by others. Equally, with the volume of data available today the accuracy of associated meta-data and sequence tags / labels is vital for ensuring the data are retrievable when broad, automated, searches are used. I suggest that a focus on quality control for additional information about each sample will make a noticeable difference to the ease with which public databases can be mined for relevant information and this exceptional resource exploited. I hope that this discussion, whilst highlighting common pitfalls, provides solutions and suggestions to guide the process of compiling data sets from online databases.

The full package vignette can be found in the supplementary files for this chapter (Appendix B).

Chapter 4

Bayesian Skyline Plots do not agree with range size changes based on Species Distribution Models for Holarctic birds

Abstract

During the Quaternary, large climate oscillations had profound impacts on the distribution, demography and diversity of species globally. Birds offer a special opportunity for studying these impacts because surveys of geographical distributions, publicly-available genetic sequence data, and the existence of species with adaptations to life in structurally different habitats, permit large-scale comparative analyses. We use Bayesian Skyline Plot (BSP) analysis of mitochondrial DNA to reconstruct profiles depicting how effective population size (N_e) may have changed over time, focussing on variation in the effect of the last de-glaciation among 102 Holarctic species. Only three species showed a decline in N_e since the Last Glacial Maximum (LGM) and seven showed no sizeable change, whilst 92 profiles revealed an increase in N_e. Using bioclimatic Species Distribution Models (SDMs), we also estimated changes in species potential range extent since the LGM. Whilst most modelled ranges also increased, we found no correlation across species between the magnitude of change in range size and change in N_e. The lack of correlation between SDM and BSP reconstructions could not be reconciled even when range shifts were considered. We suggest the lack of agreement between these measures might be linked to changes in population densities which can be independent of range changes. We caution that interpreting either SDM or BSPs independently is problematic and potentially misleading. Additionally, we found that N_e of wetland species tended to increase later than species from terrestrial habitats, possibly reflecting a delayed increase in the extent of this habitat type after the LGM.

Keywords: demographic history, mitochondrial DNA, Bayesian skyline plots, species distribution models

4.1 Introduction

The Quaternary period has been characterised by extensive cycles of glaciation and deglaciation. The legacy of these ancient large-scale climate alterations is evident today in everything from species' genetic diversity to population structure [164, 165]. Despite the profound impact that past climate changes have had on both flora and fauna, there is limited quantitative evidence on which factors determined how different species fared during these cycles, or how species responded to subsequent post-glacial climate amelioration.

One of the most widely used genetic methods for inferring demographic history is the so-called skyline plot, a family of graphical, non-parametric methods first introduced by Pybus *et al.* [51]. Grounded in the principles of Kingman's coalescent theory [53], the 'skyline framework' aims to use DNA sequence data to reconstruct a gene tree. The rate of coalescent events within the gene tree can then be used to infer how the population changed in size over time: in essence, periods of low coalescent rates imply a large population while a high density of coalescent events implies a small population. Although skyline plots have been used to reconstruct demographic histories for many species, both extant and extinct [68], and across taxa that include vertebrates [166, 167], invertebrates [67, 69], and even bacteria [70], comparative studies across many species are only now emerging [104].

Skyline plots have been used extensively to infer the response of species during the Last Glacial Maximum, and they are often paired with climatic reconstructions to infer the changes in available habitat for a given species [168, 169, 170]. One popular approach for reconstructing possible changes in available habitat for a species through time is the use of bioclimatic Species Distribution Models (SDMs) [171]. Modelling algorithms combine data on occurrences with environmental data to describe a species' current distribution and then simulate how changes in environmental variables may have influenced their range over a period of interest. The underlying logic in linking these approaches is that, assuming limited population structure and appropriate sampling, a skyline plot could, in principle, provide an indication of changes in total population size, and thus of the range occupied by a species. However, the association between effective population sizes (N_e) as reconstructed by skyline plots and species ranges is generally assumed rather than tested.

There are a number of reasons why reconstructed N_e might not be a good proxy for species ranges. Much attention has been devoted to population structure as a confounding effect, and the recommendation to counter its effects is to pool samples from multiple locations [82]. However, even with this sampling scheme, there might be a mismatch in the two quantities if mean population density, and thus N_e, was affected by climate change differently to total range extent. For example an increase in mean population density, and thus population size, might occur without a change in range, if the quality of habitat and its

carrying capacity increased without a change in its extent (Figure 4.1A) [172]. Given the positive relationship generally observed between range extent and mean local population density [173], N_e would also be expected to increase by a greater proportion than range extent under climatic amelioration. Another plausible cause of discordance between changes in N_e and range size is that, without substantial gene flow, skyline plots will mostly reconstruct the population dynamics of the sampled locations rather than the whole species [174]. Pooling samples from multiple locations can help, but it will not fully resolve the problem [82].

Another conceivable scenario that might lead to a disconnect between local N_e and range size arises during range shifts, as sampled locations, which are suitable for a species at present day, might have been only marginally suitable in the past. In other words, what is now thought of as the core area occupied by a species (i.e. where it is abundant, and sampling is more likely) might be inhabited by populations that in the past were at low densities because the local habitat was only marginally suitable. A skyline from such populations would reveal a strong increase in N_e which reflects the local amelioration of conditions for that species, irrespective of broader range changes (Figure 4.1B). A similarly confounded signal will be found in the more extreme scenario where, as a result of a sizeable range shift, the sampled populations inhabit areas which were completely unsuitable in the past, and thus have undergone a founder event after the Last Glacial Maximum (LGM). Such populations would be characterised by a steep increase in N_e as they recovered from the local bottleneck associated with the founder event (Figure 4.1C).

The extent to which changes in population density and bottlenecks related to range shifts can override the signals linked to changes in range size and the overall metapopulation size is unknown. In the current paper, I mined GenBank to compile a comprehensive dataset of publicly-available mtDNA sequence data from many species of Holarctic birds, and reconstruct their population dynamics using Bayesian skyline plots (BSP) in BEAST2 [115]. A simple prediction, based on the relative changes of habitat types as reconstructed from the pollen record [175], is that species associated with closed habitats (such as forests) should have increased since the LGM, whilst species from open and semi-closed habitats (for instance grasslands and steppes) should show a decrease. However, species have more complex niche requirements than a simple association with a broad habitat type, and a more realistic prediction is that N_e should change in line with changes in extent of the potential geographical range, such as that reconstructed by bioclimatic SDMs. I therefore reconstructed changes in modelled potential range extent between the LGM and the present, using paleoclimate reconstructions and Species Distribution Models, and investigated the relationship between N_e changes and range size changes. I acknowledge that reconstructing detailed individual ranges in the past is challenging because species may depend upon

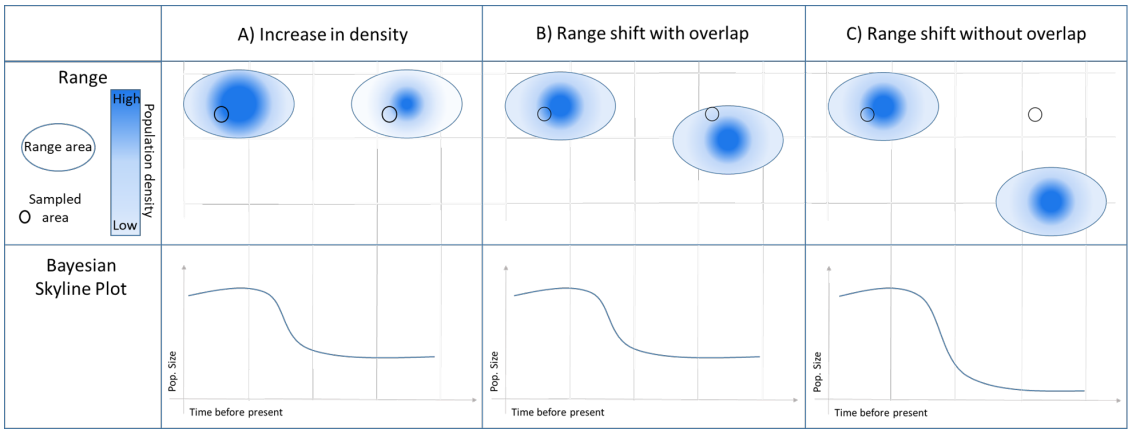


Fig. 4.1 Three scenarios that might lead to an increase in N_e without a change in range size. The top half of each panel represents a schematic map of the species range at the LGM (right) and present day (left). The density of colour within each range ellipse shows population density. Circles on the map represent the genetic sampling location. The scenarios are: A) An increase in population size without a change in range recovers an increasing BSP profile. B) Core area today was only marginally suitable in the past, range size remains the same but local amelioration has led to a strong local increase in N_e. C) Sampled populations inhabit areas outside the species range in the past. BSP recovers a steep increase in N_e associated with founder event.

habitats whose extent is not easily reconstructed (e.g. wetlands); therefore, I also compared the demographic reconstructions of species grouped by major biomes in order to investigate whether there was any consistent pattern in their response to climatic amelioration in the Holocene.

4.2 Results

4.2.1 Summary of available BSPs

Based on criteria of having a minimum of 10 individuals sequenced for either the NADH dehydrogenase subunit 2 (ND2) or cytochrome b (cytb) genes from the mitochondrial genome (mtDNA), a scan of GenBank yielded a preliminary dataset of 208 species. From these, datasets were discarded for the following reasons: insufficient haplotypes captured for demographic reconstruction ($<$ five), insufficient sequence length ($<$ 200bp), sequences across studies not from comparable sections of the gene, haplotype frequencies not published, an inappropriate sampling strategy used by the original study (e.g. non-random sampling, localised island populations) or extensive population sub-structure (see Materials and Methods for details on the criteria used to select suitable datasets). Application of these criteria

left 167 datasets for BSP analyses. All these datasets were analysed with BEAST using a Bayesian Skyline Plot and, with one exception (King Eider, *Somateria spectabilis*), they converged successfully. It is worth noting that in BEAST, I adopted a strategy of resizing the *'bGroupSizes'* parameter (see Materials and Methods), potentially constraining the level of detail recoverable in the profiles but ensuring that a large volume of variable quality datasets could be analysed using the same settings (and thus providing comparable estimates).

Data for both ND2 and cytb were available for 28 species. For 18 species, expansion times and profiles across both genes were consistent, and a single profile was then selected to illustrate the population history. Ten populations showed discordant demographic histories but, in 7 of these cases, one dataset was of appreciably lower quality (e.g. fewer samples, shorter sequences, inappropriate sampling strategy) and was removed. Three species were rejected because the two genes gave discordant profiles despite both appearing to be of comparable quality. See Supplementary Table C.1 for details of datasets dropped.

Further profiles had to be excluded as: profiles were either too deep (limit $=> 1,000,000$ years before the present, $n = 9$) or too short (limit $=< 5,000$ years before the present, $n = 4$) to be informative for the last de-glaciation; or profiles showed patterns of expansion or contraction that predated the time period of interest, Marine Isotope Stage 3 (~60 kya, $n = 18$)[176]. Note that, even among the accepted profiles, there remains great variation in depth due to the sparser and more stochastic branching patterns at the bases of the trees, which cause many profiles either to truncate or to 'flatline'. Thus, the oldest population size estimates tend to be approximations both because of the reduced information content that impacts all profiles, and the need to use the points of truncation in short profiles or flatline states as the oldest size. For the two species where multiple lineages were identified (pine grosbeak, *Pinicola enucleator*, and horned lark, *Eremophila alpestris*), two separate BSP analyses were performed and an average of the estimates was taken for downstream analysis.

Applying the above filters left 102 qualifying species BSP profiles for further analysis. These species inhabit a wide range of habitats Closed ($n = 43$), Open ($n = 17$), Semi-closed ($n = 25$), Wetlands ($n = 12$), and Other ($n = 5$); see Materials and Methods for a description of how habitats were grouped. There was no indication that species associated with particular habitats were more or less likely to be excluded ($p = 0.77$, Fisher's Exact Test, excluding the 'Other' category as it had too few species for testing). Skyline profiles encompass a wide range of shapes, variously exhibiting a single sharp point of inflection, gradual changes in size and multiple points of change. No significant differences were found in the proportion of ND2/cytb genes in each habitat type ($p = 0.78$, Fisher's Exact Test, excluding the 'Other' category), nor in the proportions of species from the Palearctic, Nearctic or Holarctic in each habitat type ($p = 0.10$, Fisher's Exact Test, excluding the 'Other' category).

4.2.2 Direction and magnitude of demographic change

Only 3 out of 102 species showed an overall decrease in N_e over time, with 7 showing no sizeable change, all other species ($n = 92$) increasing to some degree (example profiles can be seen in Supplementary Fig. C.1). The direction of change was not associated with habitat ($p = 0.457$, Fisher's Exact test, excluding the 'Other' category), nor was its magnitude (gls $p \geq 0.402$ for backbone E and $p \geq 0.386$ for backbone H, lm without phylogenetic correction $p = 0.667$; Fig. 4.2A). This result is rather extreme, but it could be the consequence of most species for which several samples are available in GenBank being relatively common and thus having thrived in the Holocene.

4.2.3 Direction and magnitude of change in extent of the potential geographical range

To investigate the plausibility of climate-driven changes in the extent of climatically suitable area contributing to the overwhelming majority of profiles showing an increase in N_e, we created individual Species Distribution Models (SDMs) with the R package *'biomod2'* [177] for the each of our 102 species. We used occurrences from the GBIF dataset, keeping samples with coordinate data accurate to 10km, removing observations outside the breeding range (defined as the area mapped by Birdlife [178] as being occupied by resident and breeding migratory populations). For each species, uncorrelated environmental variables for the present day and the LGM [25] we extracted. The dataset was then thinned to reduce spatial sorting bias [179] and randomly sampled 5 sets of pseudoabsences in the same number as presences from outside the breeding and residential areas. Models were run following four different algorithms [180] and created ensembles [38] by validating each by spatial cross-validation [181]. In the end, there were credible SDMs for 96 species; 5 species had to be excluded as there were insufficient observation points left for analysis after data thinning, and one species was rejected as its SDMs led to a present day projection much larger than the observed range (Supplementary Table C.2). For the valid SDMs, projected ranges were generated for the LGM (21 kya) and present day and the changes in range size were quantified.

As was the case for N_e changes, the majority of species showed an increase in reconstructed range extent since the LGM (76 out of 96). However, the proportion of species showing an increase in range extent was significantly smaller than the proportion with increased N_e ($p = 0.004$, data subset to the 96 species for which both analyses were available). Whilst there was variation among groups of species associated with different habitats in the magnitude of change in range extent (gls with phylogenetic correction: $p \leq 0.0001$ for all

1000 resolutions of both backbone E and backbone H, lm without phylogenetic correction $p = 0.0001$, Fig. 4.2B) and the direction of change in range extent ($p = 0.009$, Fisher's Exact Test, excluding the 'Other' category), there was no significant match between the direction of the trend in BSP and SDM reconstructions (Supplementary Fig. C.2, Fisher's Exact Test $p = 0.587$). Neither was there a significant positive correlation across species between the signed magnitudes of the changes in the two measures (Fig. 4.2C; gls with phylogenetic correction: $p \geq 0.994$ for all 1000 resolutions of backbone E and 1000 resolutions of backbone H, lm without phylogenetic correction $p = 0.994$). Furthermore, taking into account changes in the location of the range (i.e. the proportion of overlap between LGM and present range) also failed to explain the changes in N_e as reconstructed by BSP (Fig. 4.2D; gls $p \geq 0.734$ for backbone E and $p \geq 0.734$ for backbone H, lm without phylogenetic correction $p = 0.734$).

4.2.4 Timing of change

I next explored the relationship between the timing of the dominant population size change and habitat type (excluding the 'Other' category, as it was heterogeneous and only had 5 species). The timings of change in size for each population in the four habitat types are presented in Fig. 4.3A. Major size change events in wetland-associated species tended to be more recent than for species from the other three habitats (gls $p \leq 0.044$ in 1000 resolutions of backbone E and $p \leq 0.044$ in 1000 resolutions of backbone H, lm without phylogenetic correction $p = 0.044$). Similar results were obtained when I excluded species which changed less than 10% in N_e (which might have added noise) (Supplementary Fig. C.3). When using molecular evolution rates from Nabholz *et al.* [156] 'Calibration set 4' the timing of all expansions are generally consistent with a response associated with the Last Glacial Maximum (LGM). However, I note that using the rate from 'Calibration set 2' (which includes older nodes than set 4) would recover older expansion dates (data not shown); given that such dates would correspond to periods of high ice coverage they seem less likely.

For an alternative view of when the expansions occurred, I further used a multi-species index (MSI). The MSI depicts normalised changes in size averaged across species within each habitat type for each time point (Fig. 4.3B). Despite exploiting a different aspect of the BSP profile shape, mean change at each time point rather than a single mean date of maximum change, MSI profiles reveal a pattern that is strongly supportive of the previous result where wetland species expand appreciably later than species in the other three habitats.

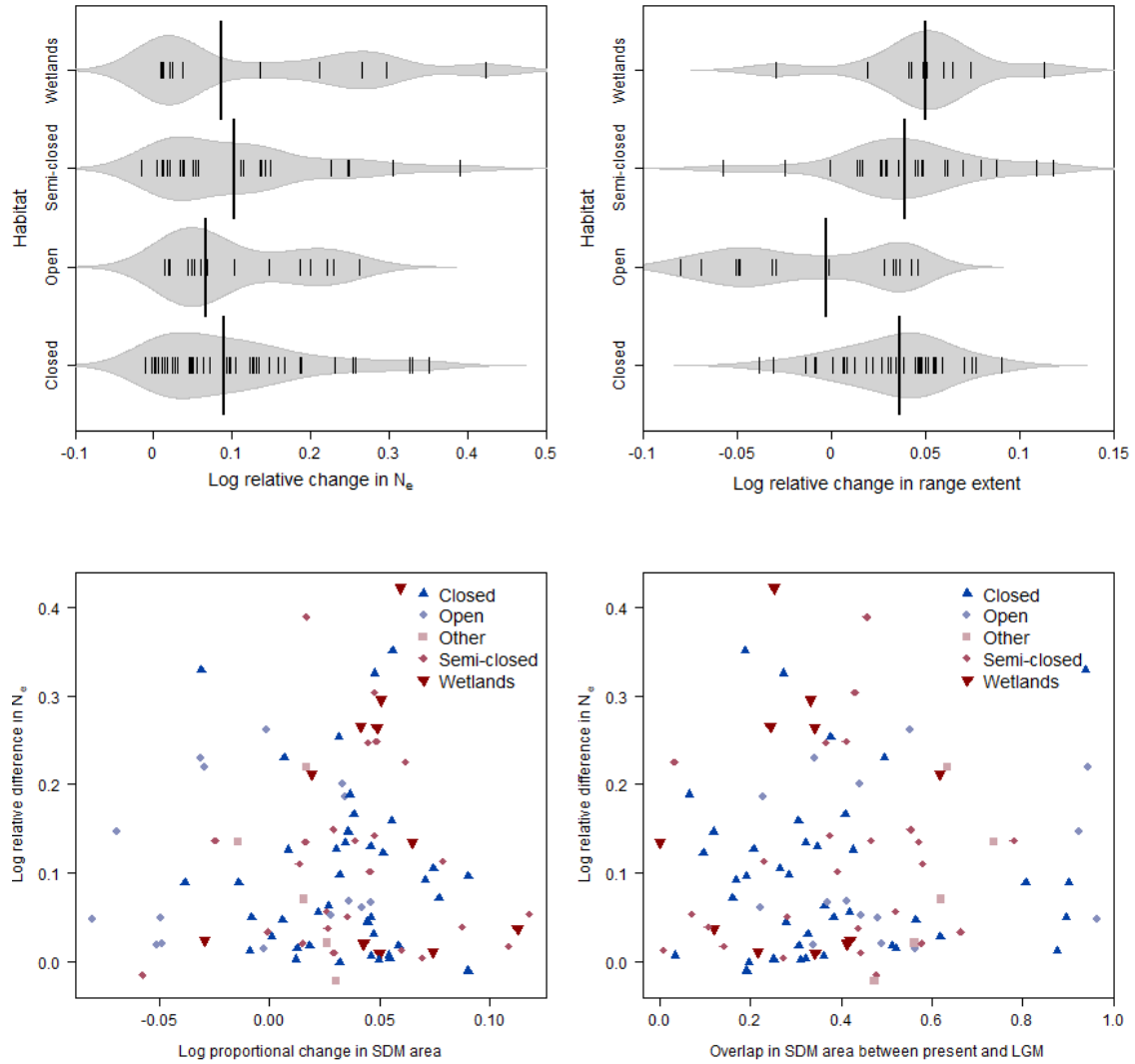


Fig. 4.2 A) Beanplot showing the log relative difference in effective population size (N_e) from 60 kya or start of the profile for species from each habitat type. Kernels represent density (i.e. frequency distribution), each small line an individual population, thick black line is median of species-specific values for the given habitat class. Numbers of species per group are; Closed ($n = 43$), Open ($n = 17$), Semi-closed ($n = 25$), Wetlands ($n = 12$). B) Beanplot showing the log relative difference in modelled range extent from 21 kya for species from each habitat type. Kernels represent density (i.e. frequency distribution), each small line an individual population, thick black line is median of species-specific values for the given habitat class. Numbers of species per group are; Closed ($n = 40$), Open ($n = 15$), Semi-closed ($n = 25$), Wetlands ($n = 11$). C) Scatterplot of log ratio of N_e from 60 kya to 5 kya in relation to the log ratio of change in size of climatically suitable area from 21 kya to the present, based upon species' individual bioclimate SDMs. D) Scatterplot of log change in N_e in relation to the proportion of the species' contemporary range that was also suitable during the LGM. In both scatter plots numbers of species per group are; Closed ($n = 40$), Open ($n = 15$), Other ($n = 5$), Semi-closed ($n = 25$), Wetlands ($n = 11$).

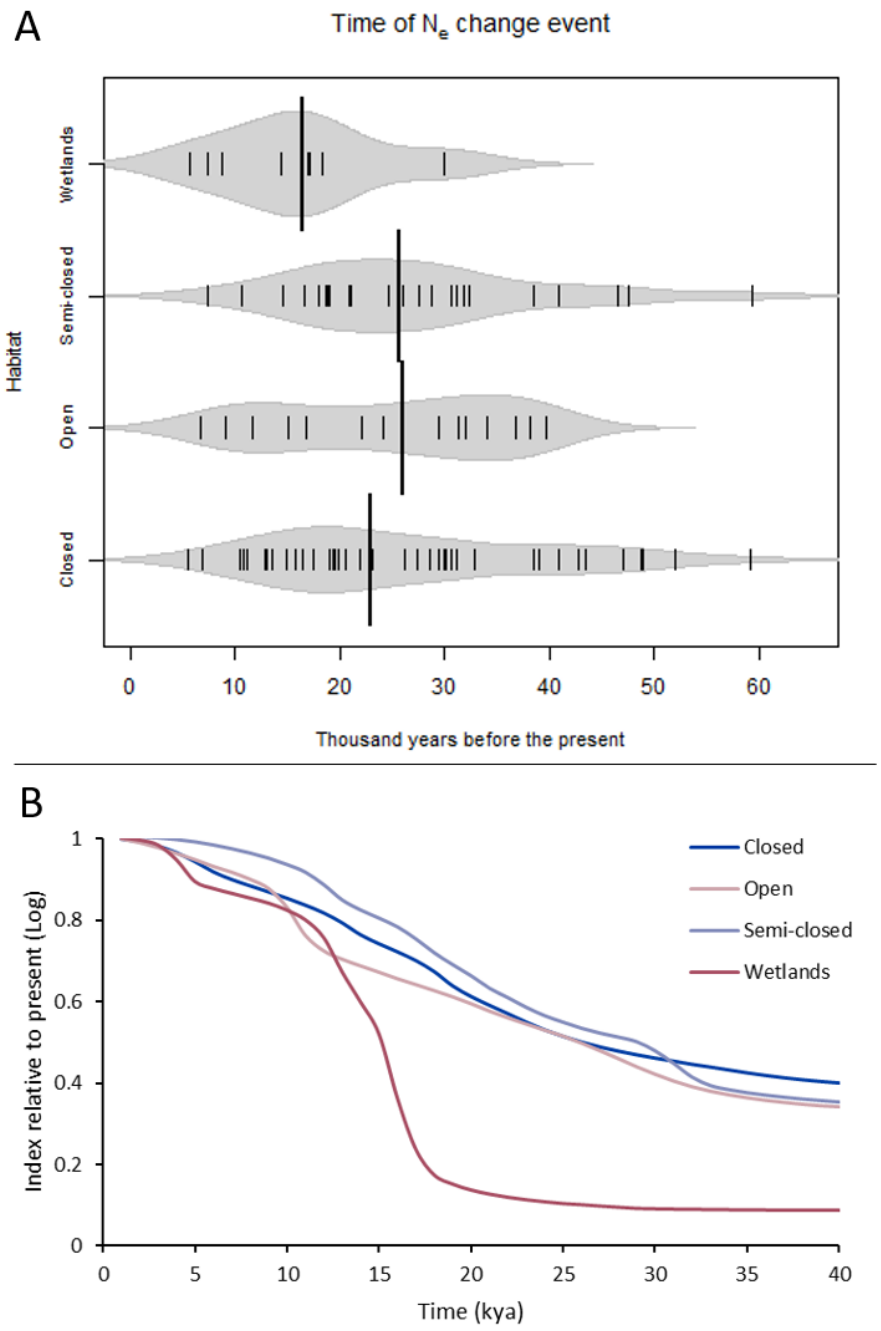


Fig. 4.3 A) Beanplot showing time of dominant effective population size change events for species from each habitat type. Kernels represent density (i.e. frequency distribution), each small line the time of an individual population's size change event (increase or decrease). Thick black line is median of species-specific change times for a given habitat class. Numbers of species per group are; Closed ($n = 43$), Open ($n = 17$), Semi-closed ($n = 25$), Wetlands ($n = 12$). B) A multi-species index (MSI) depicting normalised changed in size averaged across species within each habitat for each time point.

4.3 Discussion

I generated a large collection of mitochondrial DNA datasets from many bird species to look for evidence of habitat-associated trends in population size through time. Although variable data quality may lead to uncertainties about the magnitude of any particular change in population size I detect, the direction of change is relatively robust [57]. Out of 102 species, only three species show an overall decrease in effective population size. Changes during the last de-glaciation in the modelled extent of the geographical range also indicated increases for most species, though the proportion of increases was lower than for N_e. However, I could find no association across species between the direction or magnitude of change in N_e and habitat or range reconstructions.

Species with very large ranges at present are the likely "winners" in terms of their response to climatic change and are thus more likely to show an increase in range and population size from the LGM. Widespread species might also be more likely to have been sampled, and indeed, I found that, based on BirdLife breeding range data, the species studied here have significantly larger modern-day ranges than Holarctic species as a whole (Wilcoxon rank sum $p < 0.001$, Supplementary Fig. C.4). The majority of species we sampled also showed an increase in range size based on SDMs. However, the proportion of expanding species according to SDMs was much lower than the one observed for BSPs, thus failing to fully explain the ubiquity of expanding BSPs, and there was no statistical association between changes since the LGM as reconstructed by BSP and SDM.

Colonisation bottlenecks during a range shifts can, in principle, lead to an increase BSP irrespective of the overall change in range size, as long as migration is low enough (i.e. if the BSP captures the local dynamics in a given population/small geographic region rather than the whole range). However, if this mechanism was important, we would expect species exhibiting an increase in BSP despite a range contraction to be associated with large shifts; this is not the case (Supplementary Fig. C.5). Therefore, colonisation bottlenecks do not seem to explain the ubiquity of expanding BSPs in our dataset.

Increases in migration can also lead to an increasing BSP without any change in census population sizes. The potential role of migration in producing counter-intuitive N_e estimates when assuming panmictic populations for a whole species has received much attention recently in the context of interpreting cross-coalescence (MSMC) profiles [81]. However, it is difficult to envision a scenario where migration would increase significantly in the face of a range contraction; the effect of migration is more likely to be seen during an expansion, when previously isolated fragments are reconnected. Thus, it seems unlikely that migration can explain the ubiquity of increasing BSPs in Holarctic birds.

A final, more likely but difficult to test explanation for these results is that population densities have increased since the LGM. Thus, even for species that have experienced a range contraction, there might have been changes in local population dynamics such that the average density is higher at present. This decoupling of range extent and density makes interpretation of the SDMs and BSPs very challenging. To resolve them, there is a need to use species abundance models that explicitly predict population densities rather than presence/absence [182, 183, 184]. Whilst fitting such models is possible in principle, they have been little used because extensive population density information is rarely available for any given species. However, recent efforts such as those by the Cornell Lab of Ornithology (https://ebird.org/science/status-and-trends/)[185] have started collating such information, opening a window in better understanding the link between range size and density. My results, however, do raise a caveat in the interpretation of SMDs and BSPs for extinct species; arguably, the best strategy is to couple the two approaches, as only their combined results might provide a good overview of the fate of a species.

The timings I find for when population expansions occurred agree broadly with those of changes in climate after the LGM. Dates were based on mitochondrial mutation rates calibrated for body size and based on a calibration set that included relatively young species splits [156], and thus likely to give faster mutation rates (i.e. less affected by selection) that were appropriate for within species analysis [186, 93, 187]. It is noteable that using a calibration set that included older species splits [156] would lead to much older (well before the last glaciation), and thus less realistic changes in effective population sizes. However, I strongly caution the reader that mutation rates calibrated by bird body size, whilst the best available option for comparative analysis, are likely to be very noisy, and individual species estimates should not overly interpreted. Ideally, one would need taxon-specific mutation rates [161] which are simply not available for the number of species investigated in this study. Having said this, the fact that species from the same habitat tend to yield broadly similar profiles gives me confidence that the relative timings are likely robust, even if the absolute values still have room for improvement.

I found that changes in N_e for wetland-associated species have occurred more recently than those from terrestrial habitats. Although the significance is marginal when based on point estimates for the date of most rapidly changing size, the finding is supported by multispecies index analysis, which also reveals a pattern of later expansion among wetland species. Compared to many terrestrial habitats, wetlands tend to be less stable. Factors such as local water table levels, the amount of meltwater from retreating ice-sheets and rates of soil erosion all play into wetland habitat development and could have delayed the establishment of stable wetland habitats after the LGM. Although reconstruction of wetland environments

and the modelling of wetland recovery is difficult [188, 189, 190, 191], analysis of pollen across Eurasia shows that species associated with wetlands such as Sphagnum moss and Alder trees both exhibit much later expansions compared with terrestrial species [192, 21]. Indeed, the expansion of alder relative to other trees [192] matches closely the relatively later expansion we find for wetland versus non-wetland birds.

BSP analysis is powerful but depends on a number of assumptions that are rarely met in real data, most notably the use of a random sample of individuals drawn from a panmictic population [137, 151, 82]. Consequently, most profiles should be seen as approximations that are easy to over-interpret [57, 174]. However, increasing numbers of public domain datasets open the door for studies based on characteristics averaged across multiple profiles constructed from species or populations that share a common habitat or other trait. This averaging approach is not without its own significant challenges and the data still need stringent filtering. In this study, I constructed and inspected network diagrams for each dataset, allowing species with genetic outliers and evidence of strong population substructure to be identified and either divided or excluded. The large number of species investigated allowed me to see a clear pattern of population expansion in almost all species following the LGM, irrespective of their range dynamics, and a tendency for the expansion to occur later in wetland species. The near-ubiquitous signal of expansion suggests a decoupling of range size and local densities, implying a need for carefully interpretation of BSP to describe species-wide responses.

4.4 Materials and Methods

4.4.1 Raw genetic data

I assembled two databases, one for NADH dehydrogenase subunit 2 (ND2) and one for cytochrome b (cytb); these two genes are among the most frequently uploaded avian mtDNA loci in GenBank. First, summary information on all available avian ND2 and cytb sequences in GenBank was collated using a custom R script, it was then screened for Holarctic species using the list of Voous *et al.* [193]. I only retained species with more than 10 accession for either gene; when a species had sufficient data for both ND2 and cytb, sequences for each gene were extracted and handled as distinct datasets.

4.4.2 Alignment

Sequence data for each species / gene combination come from multiple independent studies and often differ in the gene region they analyse. Comparable regions were found by aligning

sequences in MEGA (version 7.0; [194] using the programme ClustalW [195]. Sequence data for each taxon were then trimmed to the longest common section between all samples. If inclusion of a single sequence required the loss of $>$ 200bp from more than 50% of the other sequences, that sample was excluded. Furthermore, all positions containing insertions, deletions or sequencing ambiguities were removed. When studies uploaded only one copy of each haplotype, I used haplotype frequencies from the associated publications to generate the appropriate number of copies in our database. Publications that lacked haplotype frequency were excluded from the analysis. After frequency correction the available sample sizes varied from 11 to 453 sequences per species, with lengths from 236 base pairs (bp) to 1137 bp.

4.4.3 Median Joining Networks

For each species / gene combination, I built a median joining network (MJNs) in POPART [196]. If the MJN contained long branches, defined as 30 or more nucleotide substitutions on a single branch, the sampling location for that species was reviewed because such long branches are indicative of profound population substructure. If clear geographical separation or grouping was found, the data were divided as appropriate and treated as discrete datasets. Single samples with $>$ 30 mutations on a branch were considered extreme outliers and dropped from alignments.

4.4.4 Mutation rate

Recent work [156] proposes that body-mass can be used to inform more accurate calculations of taxon-specific substitution rates and provides a correction factor for variation in rates according to body mass as well as major mtDNA loci. I created dataset-specific molecular evolution rates using the body mass / gene correction factors from Nabholz *et al.* [156] 'Calibration set 4' (3rd codon position), as it includes younger species splits that should lead to estimates more appropriate for within the within species dynamics investigated in this chapter. Due to the uncertainty surrounding mutation rates, analyses based on 'Calibration set 2' were also run (data not shown). Body mass data was taken from Dunning *et al.* [197].

4.4.5 BSP analysis

Whilst there now exist a range of related skyline plot methods (see [71] for a technical review) I focus on BSPs [60]. The relative simplicity of this approach, and its inherent robustness, makes it particularly suitable for the heterogenous quality of datasets investigated in this chapter. For each dataset, BSP analyses were implemented in BEAST2 [115, 66] using a

strict clock with a taxon-specific body mass / gene mutation rates, run lengths of 300 million steps sampled every 30,000 steps, with the first 10% discarded as burn-in. The integrated Bayesian application *'bModelTest'* was used to select the most appropriate site model and parameters for individual analyses [198] and *'bGroupSizes'* was set to 3. All other parameters were left as default. Each analysis was run twice and convergence verified by both a visual inspection of MCMC trace output in Tracer v1.6 and confirming that the effective sample size (ESS) values exceed 200 [114]. Demographic reconstructions were then summarised in Tracer v1.6 (http://tree.bio.ed.ac.uk/software/tracer/) with *'Number of bins'* set to 500, and plotted in R.

4.4.6 Inclusion criteria

Where data were available for both ND2 and cytb BSP profiles were compared along with summary statistics on each dataset. When profiles were in agreement, the best supported dataset was retained, e.g. largest sample size, longest sequences, to represent that species' history. If profiles were not concordant but there was a clear disparity in the quality of the datasets, I again kept the profile from the better dataset. If the profiles did not show similar trends, and there was no difference in the data quality, I conservatively rejected both profiles. For inclusion in further analysis, profiles needed to have a history deeper than 5 thousand years ago (kya) but shallower than 1 million years ago, and also have recovered a change event (increase and / or decrease) dated within the last 60 kya.

4.4.7 Habitat classification

The species considered here are associated with a wide range of habitat types, especially in terms of the predominant vegetation. Given the need to use a small number of habitat categories, no classification system can be perfect. I used the expert ornithological opinion of one of us (REG) to classify each species according to the major habitat with which each species is currently associated, based upon descriptions of their natural habitats in a standard work [178]. After initial data quality filtering 138 species were available to be classified this way. The habitat classes selected were Closed (forests), Semi-closed (shrubland and open woodlands), Open (grassland, montane and steppe), and Wetlands (freshwater wetlands). Some species could not be placed in one of these classes and there were other classes (e.g. Rivers) with five or fewer species. These exceptions were grouped into a category referred to as 'Other'.

4.4.8 Timing of expansion

Identifying a population's point of expansion from a BSP profile proved difficult given the wide range of shapes present: some populations changed little or very gradually in size, while others showed sharp and / or multiple points of inflection. I chose to identify inflection points using a custom algorithm, however, as timings could be confounded by the capture of local optima I also used visual inspection of the plotted data and a custom R script to review and extract exact timings. The 'algorithmic' method used a moving window approach, repeated for five different window sizes. In each window a linear regression was fitted, and its equation used to remove any slope, before fitting a second order polynomial and recording the difference in y-axis value between the mid-point and the average start and end of each window. This method usually identifies one point of maximum turn (increase or decrease) but the visual inspection of each plot allowed us to capture any additional size change events that were present.

4.4.9 Size change

To compare the magnitude of estimated population expansions within the period of interest, the relative population size change between 60 kya and 5 kya was calculated. N_e at 60 kya and 5 kya was interpolated for each BSP profile and where profiles were shorter than 60 kya the population was assumed to be the size at the start of profile. Given the uncertainty of molecular methods I preferred to use size estimates from 60 kya / the earliest possible point instead of estimating the size at the height of the LGM (21 kya) where it is more likely the analysis would catch profiles already undergoing demographic change as a result of climatic events.

I also created a multispecies index (MSI) to further explore the changes in N_e. MSI is a form of average profile based on the average ratio of estimated population size between adjacent time points. Specifically, for each pair of adjacent time points I calculate the geometric mean of the ratios of all species for which data exist. A profile for each broad habitat grouping was then constructed based on these ratios, working back from the present which is assigned a value of 1.

4.4.10 Phylogenetic correction

The species included in this study are not phylogenetically independent, so all analyses included a phylogenetic correction based on the most complete molecular phylogeny of extant birds (www.birdtree.org, [199]). Jetz *et al.* [199] present two phylogenetic backbones,

the Ericson and Hackett backbones ('backbone E' and 'backbone H' from now on), and, as we considered both to be equiprobable, all analyses were repeated with both. For each backbone, we generated 1000 trees, randomly resolving polytomies in each, and repeated all analyses for each tree with a "pgls" phylogenetic correction from the *Caper* package in R. For all analyses, we also provide the results based Ordinary Least Squares (i.e. without phylogenetic correction)..

4.4.11 Species Distribution Models

The whole pipeline is provided as a commented R script in the supplementary materials. C.

Present day and LGM paleoclimate reconstructions:

In order to identify areas suitable for species through time, high-resolution climate data from the past is needed. We used a 0.5° resolution dataset for 19 bioclimatic variables; Net Primary productivity (NPP), Leaf Area Index (LAI) and all the BioClim variables [24] with the exclusion of BIO2 and BIO3; covering the last 21,000 years in 1,000 year time steps [25]. This dataset was originally constructed from a combination of HadCM3 climate simulations of the last 120,000 years, high-resolution HadAM3H simulations of the last 21,000 years, and empirical present-day data. The data were down-scaled and bias-corrected using the Delta Method [200].

Species data preparation:

Species occurrences were downloaded from the GBIF database (https://www.gbif.org) without any preliminary filtering (download links are available in Supplementary Table C.2). Occurrences were then filtered based on the accuracy of the coordinates (maximum error: 10 km), keeping only observations within the breeding and resident geographical ranges from Birdlife [178]. The occurrences were then regridded based on the palaeoclimatic reconstructions ($0.5°x0.5°$) and, as the method used works on presence / absence and not frequency, only one presence per grid cell was kept.

For each species, this cleaned dataset of presences was then used to select a subset of bioclimatic variables from the 19 variables available in the paleoclimatic reconstructions [25]. In order to avoid using highly correlated variables, which may increase noise in the data [37], a correlation matrix was constructed between the variables associated with each presence. Where two values were highly correlated, the variable with the lowest overall correlation across the matrix was retained and this way a set of uncorrelated variables (threshold $= 0.7$) were selected.

Opportunistic observations, such as those collected in the GBIF database, tend to have geographic biases in sampling effort. In order to reduce the risk of geographic sample bias affecting the SDMs, the dataset was thinned using the R package *spThin* [201], a minimum distance of 70 km was enforced between observations. Given the random nature of removing nearest-neighbour data points, this process was repeated 100 times ('rep' = 100) but in order to keep as much information as possible the result which kept the maximum number of observations after thinning was retained for downstream analysis.

Species Distribution Model fitting:

Species Distribution Modelling was performed using the R package *biomod2* [177] for all species with more than 10 occurrences after filtering and thinning [202]. The thinned dataset was used as presences, the whole region (i.e. the land mass of Eurasia or North America) as background, and then the same number of pseudo-absences as presences were randomly drawn from land outside the BirdLife resident and breeding masks 5 times, creating 5 independent datasets for further analysis. I found that drawing pseudo-absences in this way, from outside the masks, was the most effective strategy for retrieving SDMs consistent with the best estimates of a species' modern-day range. By confirming that the estimated distributions recovered for the modern day were in accord with the BirdLife range predictions we were confident that the modelled niche being projected into the past was with as accurate as possible (an example using the *Passer domesticus* dataset is given in Supplementary Fig. C.6).

Following Bagchi *et al.* [180], models were run independently for each of the five pseudoabsence datasets using four different algorithms: generalised linear models (GLM), generalized boosting method (GBM, in the mentioned reference defined as "boosted regression tree"), generalised additive models (GAM) and random forest. Model evaluation was performed by spatial cross validation [181], i.e. splitting the dataset (both presences and all five runs of pseudoabsences) based on latitudinal bands in America (Supplementary Fig. C.7) and longitudinal bands in Eurasia (Supplementary Fig. C.8) with the R package *BlockCV* [203], and using 4/5 of the splits to calibrate the model and the remaining 1/5 to evaluate it. Latitudinal or longitudinal bands were chosen simply based on the shape of the continent. A data split cannot be used for evaluation if it contains only absences. For this reason, given the great variety of distribution of the species analysed, I decided to maximise the probability of having at least some presences in all data splits by creating 15 spatial blocks encompassing the whole region of interest, either North America (East-West bands) or Europe (North-South bands). Each block was given an ID, numbered sequentially 1-5, the 15 blocks were then assembled into five working data splits grouped by the assigned ID numbers.

The models were run five times (once for each pseudoabsence run) for each of the four mentioned algorithms, using in turn four of the five defined data splits to calibrate and one to evaluate based on TSS (threshold $= 0.7$).

A full ensemble, combining all pseudoabsences sets and algorithms [38], was then built using only models with TSS > 0.7 averaged through four different statistics: mean, median, committee average and weighted mean. The statistic showing the highest TSS was then projected to either Eurasia or Northern America considering both the present-day climate and the palaeoclimatic reconstruction for the LGM (defined as 21 kya).

Range change and overlap:

Finally, the binary projection was used to estimate the climatically suitable area (in square kilometres) for each species both now and in the LGM. In order to do so, I first re-projected the rasters to the Eckert IV equal-area pseudocylindrical projection setting the grid size to 50x50 km, and then multiplied the number of cells occupied in each period, and their overlap, by the cell area (2500 km^2).

4.4.12 Range size comparison between sample species and all Holarctic species

I used BirdLife data on the Extent of Occurrence (in km^2) to define range size of ~9000 Holarctic bird species, defined as those with a mean range latitude of above 20°N. We then compared this full dataset to a subset of range sizes for the 102 species included in our study using a Wilcoxon rank sum test.

Chapter 5

Exploring the Demographic Signals that can be Recovered from Populations During Range Shifts Using a Spatially Explicit Reconstruction of Post-glacial Recolonization in North American Yellow Warblers (*Setophaga petechia*).

Abstract

Changes in population demographics, such as population size, leave signals in the genetic variation of individuals alive today. Therefore, sequence data from contemporary individuals, combined with population genetics theory, can be used to explore patterns of long-term demographic changes going tens, or even hundreds, of thousand years back in time. Bayesian skyline plot (BSP) methods have become a popular way to explore the changes in population demographics from modern genetic data. However, patterns of population size change from skyline plots are often heavily over-interpreted in ways that I have highlighted in earlier chapters. In the current chapter, I explicitly reconstruct the post-glacial recolonization of the yellow warbler (*Setophaga petechia*) using a spatial model fitted to real genetic data. By modelling mitochondrial DNA samples from three different regions of the species' contemporary range, I am able to explore the effect of historic range dynamics on recoverable demographic patterns found using BSPs. Unfortunately, despite a well-fitting model, the complexity of the system appears to preclude accurate demographic profile recovery. Thus, for the yellow warbler, the unique demographic profiles of each region are not recovered in the BSP and I am unable to use this system to provide any insight into the features and dynamics of real systems that may confound BSP analysis. However, though not capturing the 'true' population history, the profiles of population size change constructed could appear to be informative. Once again, I have shown that it can be problematic to interpret BSPs at face and I encourage future studies investigating the history of population size to employ multiple lines of evidence wherever possible.

Keywords: demographic history, spatial model, mitochondrial DNA, Bayesian skyline plots

5.1 Introduction

The current unprecedented loss of global biodiversity and the accepted understanding that anthropogenic stressors are, in a large part, responsible for this has sparked great concern in the scientific community. In turn, this concern has prompted a lot of work focusing on understanding how species are coping in today's rapidly changing environment and, indeed, how these species are likely to cope in the future. However, modern trends can only offer insight into the effects of environmental change over a short period. Understanding the dynamics of how a population operated and altered during major climatic changes pre-history can help our understanding of how species might respond to future fluctuations [204].

Changes in population demographics, such as population size, leave recognisable signals in the genetic variation of individuals alive today [46]. Therefore, sequence data from contemporary individuals, combined with population genetics theory, can be used to explore patterns of long-term demographic changes going tens, or even hundreds, of thousand years back in time. When scaled to 'real time', it is possible to integrate these historical population sizes with other ecological variables in order to explore the influence of factors such as past climate and community structure in more detail.

Skyline plots [51], or versions thereof [205, 60, 65, 63, 66], have become a popular way to reconstruct the changes in population demographics from modern genetic data. This approach, which is grounded in the principles of the coalescent theory, estimates effective population sizes (N_e) based on the density of coalescent branches in a gene tree moving backwards through time. To date, the family of skyline methods have been applied to a huge range of both extinct and extant species encompassing plants, viruses, invertebrates, and vertebrates [206, 207, 67, 208, 209]. Indeed, the approach has proven extremely useful, allowing the scientific community to resolve information about the history of many species that would otherwise have been inaccessible [68, 210].

Patterns recovered from skyline plots, which reconstruct a population's demographic history, are often considered to be linked to other key population features such as range dynamics [168, 169, 170]. However, work presented in earlier chapters has demonstrated that this isn't always the case and that it is vital demographic profiles are considered in a wider context. By explicitly reconstructing the post-glacial recolonization of a specific, extant, species it is possible to create an opportunity to explore the effect of historical range dynamics on recoverable demographic patterns found using Bayesian skyline plot (BSP) methods. Modelling samples from different regions of the species' contemporary range will allow me to investigate if underlying range dynamics are driving recovered patterns and if they may even cause misleading profiles to be reconstructed. It is important that we build on

our understanding of when and where demographic reconstruction approaches work reliably, what features might be driving them, and where they may struggle.

I chose to look at the North American yellow warbler (*Setophaga petechia*) a small, riparian, migratory passerine. Over the last 50,000 years the area of North America habitable for the yellow warbler changed considerably. During the last glacial maximum (LGM, ~21 thousand years ago (kya)), large swathes of the region were covered in ice and the majority of North America species were excluded from the north. Yet, as the continent experienced climate amelioration and new areas of habitat became available re-expansion and colonisation events have meant that, today, this common species is widely distributed across the continent. However, despite its large and well-connected contemporary range the yellow warbler has recorded a declining trend in the North American Breeding Bird Survey between 1966-2015. This declining trend in population health has triggered several studies looking into the species ability to cope in the face of a rapidly changing climate [211]. One such study was the work of Bay *et al.* [212] who built RAD-seq data from individuals sampled across the species' range in order to explore potential population trends in response to future climate scenarios. This RAD-seq data was made available on GenBank.

Whilst large datasets built from nuclear DNA markers are becoming more prevalent as a consequence of the falling costs of genetic sequencing, for the moment, these datasets are still more limited than for mtDNA. Due in large part to its simple, haploid nature, and the ease of sequencing it, mtDNA has been the 'go to' loci for genetics projects for decades. As a result, today there exists a huge volume of mtDNA data from diverse species across the globe and I therefore chose to focus on mtDNA sequence data.

In order to investigate the link between skyline plots and species range dynamics in detail I will initially explore what genetic patterns are found in the North American yellow warbler population today using RAD-seq data from Bay *et al.* [212]. I will then fit a spatially explicit model of population growth and expansion that accounts for climatic change to this empirical dataset. Using parameter values that represent a set of realistic expansion dynamics able to capture patterns found in the RAD-seq data, I shall generate a series of simulated mtDNA sequences. Theses simulated sequences will then be used to construct a series of BSPs with known population histories. By modelling mtDNA samples from different area of the species' contemporary range with distinct histories, I aim to investigate the impact of underlying range dynamics on the patterns and profiles recovered by skyline methods. Hopefully, this approach will also allow me to asses if certain demographic histories are prone to cause misleading profiles to be reconstructed.

5.2 Materials and Methods

5.2.1 Raw genetic data

First, RAD sequence data for North American yellow warblers (*Setophaga petechia*) from 21 populations [212] were downloaded from the NCBI Sequence Read Archive (SRA). From the 269 accession associated with the Bay *et al.* paper I chose to focus on only the individuals included in the original analysis ($n = 223$), individuals for which full information about their breeding population was available. A further 22 samples were dropped as the file sizes were under 75MB and, therefore, were likely to have low coverage. One final exclusion was made, GenBank accession number SRR6366039, as the sample was found to be an outlier with a measure of diversity higher than the range of all other samples, despite comparable levels of coverage and number of sites. This left 200 samples for further analysis.

These individuals were sampled from across the modern population range, providing a good overview of the population genetics of this species. I started by exploring the geographic patterns of this dataset, to establish the features that any mechanistic model needed to capture. To do this I investigated a range of features discussed below.

5.2.2 Genotype-free estimates of diversity

RAD-seq methods are known to create specific biases in estimated allele frequencies, potentially affecting downstream analysis of the data [213]. Using allele frequencies derived directly from the sequence data in a genotype-free method has been shown to account for RAD-seq specific issues, improving population genetic inferences [213]. Therefore, I used the programme Analyses of Next-Generation Sequencing Data (ANGSD) [214, 215] to infer genotype likelihoods directly from aligned BAM files. Pairwise π (the average number of differences between two sequences, normalised by the number of available positions) were then calculated from all pairs of individuals. Genetic differentiation among pairs of populations was quantified as Fst, which I calculated from the estimates of mean π for pairs of individuals that belong to the same population (π_{within}) and different populations ($\pi_{between}$) using the equation from [121]:

$$Fst = (\pi_{between} - \pi_{within})/\pi_{between}$$

5.2.3 Isolation by distance

Isolation by distance (IBD) was quantified as the relationship between linearised Fst [Fst/(1 – Fst)] and geographic distance (computed as the great circle distance), tested with a Mantel's

test [216]. When multiple locations were sampled for one population, the mean value of latitude and longitude was used.

5.2.4 Isolation by resistance

Even though birds can easily cross most terrains, large mountains can play a role in determining their movement (and thus population connectivity). I explored the impact of topography on gene flow between populations using electrical circuit-theory [217][23] implemented in CIRCUITSCAPE [218]. Altitude information for the globe was downloaded from the WORLDCLIM database and then cropped to focus on North America, this map was then re-projected using an equidistant conic projection. The altitude raster map was then turned into a resistance surface, with resistance in each cell being a function of altitude according to the function $e^{(a \cdot elevation)}$, where the exponent a determines the cost of crossing cells of high altitude.

An isolation by resistance (IBR) model then tested all possible routes between populations, measuring the ease with which genes can flow between different pairs of cells in landscape. The model was fitted against linearised Fst (as detailed for the IBD tests). The role of altitude was explored by testing the impact of different values of coefficient a (for values of 0, i.e. no effect, 0.0001, 0.0002, 0.0003, 0.0004, 0.0005, 0.001, 0.002, 0.003), thus altering the effective resistance and so the permeability of this landscape feature to dispersal.

5.2.5 Climate Informed Spatial Genetic Models

I fitted a Climatically Informed Spatial Genetic Model (CISGeM) to the yellow warbler genetic dataset. In this modelling framework, population sizes and migration are governed by simple demographic rules that respond to changing climatic conditions (as reconstructed from paleoclimate models, see section Demography for details).

Species Distribution Modelling

The range and population size of a species alters in time and space according to fluctuations in resources and environmental conditions. In order to build the spatially explicit model it was first necessary to reconstruct how populations ranges and demographics may have changed during the period of time I am focusing on. This is done using Species Distribution Modelling (SDM). For the yellow warbler population an SDM analysis was undertaken using an R [219] pipeline, available as a vignette in the supplementary material for this chapter (Appendix D).

SDM paleoclimate reconstruction: Climate data for North America were drawn from a 0.5°resolution dataset for 19 bioclimatic variables; Net Primary productivity (NPP), Leaf Area Index (LAI) and all the BioClim variables [24] with the exclusion of BIO2 and BIO3; covering the last 21,000 years in 1,000 year time steps [25]. This dataset was originally constructed from a combination of HadCM3 climate simulations of the last 120,000 years, high-resolution HadAM3H simulations of the last 21,000 years, and empirical present-day data. The data had been downscaled and bias-corrected using the Delta Method [200]. Climatic variables through time were then used as input data to inform the SDM.

SDM data preparation: Species occurrences data were initially downloaded from the GBIF database (https://www.gbif.org), the original downloads are available at the following DOI: S. petechia 10.15468/dl.jfkwcg (GBIF.org). As a first step, occurrences were filtered based on the attributed accuracy of the coordinates (maximum error: 1 km), any points that fell outside the breeding and resident geographical ranges as estimated from Birdlife [178] were also removed as likely errors. Remaining occurrences were then matched to the grid used for the palaeoclimatic reconstructions (0.5 x 0.5 degrees) and, in order to generate the presence / absence data needed for the SDM, only one observation was kept per grid cell.

A subset of the 19 available bioclimatic variables that were informative but uncorrelated then needed to be selected. 1,000 points were randomly selected from the cleaned dataset to act as a baseline from the whole of the North American landmass. In order to define which variables have the most influence on the species distribution a beanplot of the density distribution was created for each variable. These plots compared the shape of the density distribution from the 1,000 baseline points to the density distribution of cells where the Yellow Warbler has been observed (presences). If the shape of plot for both the baseline and the presences of a single variable were similar, then it could be assumed that this variable was not informative for the distribution of the species; based on that variable alone the whole region could be suitable for the species. However, variables with mismatched profiles, e.g. different tails or peaks in the beanplot distributions, were assumed to be informative and as such were retained. In order to avoid using highly correlated variables, which may increase noise in the data, the correlation matrix between each of the retained variables (presence points only) were then calculated. Where two values were found to be highly correlated, the variable with the lowest overall correlation across the matrix was retained and this way a set of a uncorrelated variables (threshold $= 0.7$) were selected.

Presence data from opportunistic databases, such as GBIF, suffers from geographic biases due to heterogeneous effort. In order to reduce this bias, the dataset was thinned with the R package *spThin* [201], enforcing a minimum distance of 70km between points. Given the random nature of removing nearest-neighbour data points this process was repeated

100 times ('rep' = 100), and, in order to keep as much information as possible, the result which kept the maximum number of observations after thinning was retained for use in later analyses.

SDM modelling: The SDM was built with the R package *biomod2* [177]. The thinned dataset was used as presences whilst the landmass of North America was considered as background. The same number of pseudo-absences as presences were then randomly drawn from land outside the BirdLife resident and breeding masks 5 times, creating 5 independent datasets for analysis. Following Bagchi et al. [180], models were run independently for each of the five pseudoabsence datasets using the four different algorithms: generalised linear models (GLM), generalized boosting method (GBM), generalised additive models (GAM), and random forest.

Model evaluation was performed by spatial cross validation, using 4/5 of the data to train the algorithm and the remaining 1/5 to test it. Initially, both presences and the five pseudoabsences runs were subdivided in 14 latitudinal bands using the R package *BlockCV* [203]. Each band was numbered sequentially 1-5 and then the bands were put together into five working data splits grouped by band ID (1-5). This was performed to maximise the probability of having at least some presences in all five data splits, because a data split cannot be used for evaluation if it contains only absences. Each of the four models (GLM, GBM, GAM, and random forest) were then run five times (once for each pseudoabsence run), using in turn four of the five defined data splits to calibrate and one to evaluate based on TSS (threshold = 0.7).

Finally a full ensemble combining all algorithms and pseudoabsences runs [38] was created, using only models with TSS > 0.7, averaged using four different statistics: mean, median, committee average and weighted mean. The mean, the statistic showing the highest TSS, was then used to predict the probability of occurrence in each cell, and it was projected for all available time slices from the present to 50 thousand years ago.

Demography

Within a CISGeM model, the world is divided into a lattice of hexagonal cells ~100km wide. Each cell represents a deme and, within each deme, the population is allowed to grow (at rate r), until local carrying capacity (K), which is determined by SDM projections, has been reached. K is a function of the probability of the deme being inhabited according the SDM projection, using the function $allometric\_scaling \times e^{(allometric\_exponent \cdot SDM\_prob)}$. Demes at carrying capacity can send colonists into empty neighbouring demes at a rate cK, whilst migration between neighbouring occupied demes occurs at rate $mN_{\min}$ where $N_{\min}$ is

population size of the smaller deme (thus resulting in symmetrical migration between the two demes). See Figure 5.1 for a schematic representation.

For each model, the system is initialised by seeding a single deme in the lattice with a population of carrying capacity K. The exact value of K is drawn from the model priors each time. The global population dynamics are then simulated forward to the present day with the number of individuals in each cell, as well as the number of migrants and colonisers moving from one cell to another, being recorded for each generation.

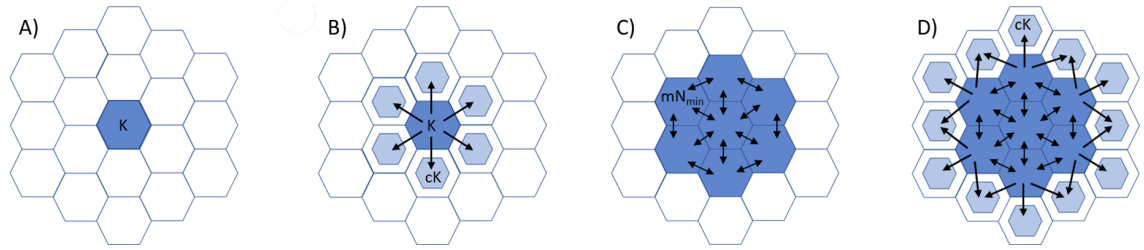


Fig. 5.1 Adapted from Eriksson *et al.* 2012 [220]. A schematic of the spatial model used in this study. A) An initial deme is seeded with K individuals from an ancestral population of K_0. B) cK colonists will move to each adjacent deme, and they grow at rate r, eventually reach the carrying capacity (K) for that deme. C) Adjacent occupied demes, once they reach carrying capacity, exchange migrants at a rate of mN_{min} where N_{min} is population size of the smallest deme. D) They also send out cK colonists to any adjacent suitable but empty deme.

Fitting a genealogy to the demography

Once a global population demography has been constructed, gene genealogies are simulated. This process is dependant on the population dynamics recorded in the demography stage and assumes local random mating according to the Wright-Fisher dynamic. From the present, ancestral lines of sampled individuals are tracked back through the generations, recording which deme each line belongs to. Every generation, the lines are randomly assigned to a gamete from the individuals within it's present cell. If the assigned individual is a migrant or coloniser, the line moves to the cell of origin for that individual before 'mating'. Whenever two lines are assigned to the same parental gamete, this is recorded as a coalescent event, and the two lines merge into a single line representing their common ancestor. This process is repeated until all the lineages have met, reaching the common ancestor of the whole sample. If multiple lineages are still present when the model reaches the generation and deme from which the demography was initialised, the lines enter a single ancestral population (K_0) until sufficient additional coalescent events have occurred for the gene tree to close.

Model fitting with ABC

Parameter space was explored with a Monte Carlo sweep in which demographic parameters were randomly sampled from flat prior ranges: r [0.01,1], c [0.02,0.166], m [0.0001,0.05] on a $\log_{10}$ scale, *allometric scaling factor* [100,5000] on a $\log_{10}$ scale, and *allometric scaling exponent* [0.1,1]. I used a mutation rate of 1.5 x10^{-9} μ/Site/Year [221].

Model fit was done within an Approximate Bayesian Computation (ABC) framework using the results of the MC sweep. To compute summary statistics, I clustered populations into three groups representing the West, Central, and East regions of the North American continent, and then computed the mean pairwise π for populations within each group and between each pair of groups, giving us a total of 6 summary statistics. In order to make the modelling computationally feasible I investigated how many samples were needed to get a good estimate of π for each population (Supplementary Fig. D.1). This analysis showed that five diploid individuals, or ten chromosomes, provided a reasonable compromise for noise. I therefore re-computed all estimates of pairwise π with only five individuals per population and checked that the estimates were consistent with the values from the full dataset, as used for the earlier pattern analysis (Supplementary Fig. D.2). For computation efficiency of the model, all future analysis will be based on this subset of the data. ABC was performed using a linear model with the R package *abc* [222]. After ABC, the top three simulations were retained and the sets of parameter values for each were used to simulate mtDNA.

5.2.6 Simulated mitochondrial DNA

Sampled populations

Three populations were chosen to represent different population histories. According to our SDMs and the demographies recovered from CISGeM, Manitoba, CA, has been comparatively recently colonised as the ice sheets retreated, whilst Oregon, USA, was colonised very early on and has maintained a stable population for the duration of the simulation. On the other hand, although Pennsylvania, USA, was also colonised relatively early, this population experienced a bottleneck event during the ice age (~21 kya) and has recovered its population size from then. For each population, I simulated a sample size of 90 mtDNA sequences (dividing the N_e obtained for nuclear markers by 4). In addition to the three separate populations, a sample set that encompassed 30 individuals from each of the three populations was also complied. This is designed to mimic a ‘full range’ sampling strategy.

Sequence simulation

Calculating accurate mutation rates is not trivial, however, it has been suggested that body-mass can be used to estimate species specific mutation rates. Using body-mass data from Dunning *et al.* and the 'Calibration set 4' (3rd codon) rates I calculated the yellow warbler mutation rate and then corrected for the cytb gene specific rate. Cytb is one of the most commonly sequenced mtDNA genes and provides a good representative rate. The final value was 1.88401 x10^{-8} μ/Site/Year.

For each of the three selected populations (Manitoba, CA; Oregon, USA; Pennsylvania, USA), sequences of 1000 base pairs and 16,000 base pairs of mitochondrial DNA (mtDNA) were generated for 90 individuals. These data were generated three times, each time using different fixed parameter values drawn from one of the three best fitting models. The outputted 'treeseq' objects were converted to a VCF and the reported mutations were then inserted into a skeleton fasta format file containing an unmutated 'neutral' reference sequence.

5.2.7 Bayesian Skyline Plots

The mtDNA sequence data was analysed using the Bayesian skyline plot (BSP) method implemented in the software package BEAST v 2.6.0 [115]. For each dataset sequences were loaded into the Bayesian Evolutionary Analysis Utility tool (BEAUti2) in FASTA format. In BEAUti, the substitution rate was set to match that used to generate the data (1.88401 x10^{-8} μ/Site/Year), the substitution model used was Jukes-Cantor (JC69), *'Coalescent Bayesian Skyline'* was selected under the *'Priors'* tab, and all other parameter and model settings were left as default. Every dataset was run twice to ensure repeatability.

Run outputs were analysed using Tracer (v. 1.7.1) [160]. Each run was set for a chain length of 300 million steps, sampled every 30,000 generations, and the first 10% was considered as burn-in. Convergence was confirmed by plotting the two MCMC chain traces together as well as checking that effective sample sizes (ESS) exceeded 200 for both runs. The BSP profiles were then built in Tracer using 500 bins and exported as a data table to be plotted in R (v. 3.5.3).

5.3 Results

5.3.1 Analysis of geographic patterns from empirical data

Firstly, I wanted to describe the patterns found within the available North American yellow warbler genetic data. With an initial exploration of the data, I confirm the strong pattern

of isolation by distance (IBD) previously found in this data [212] indicating some level of restriction to gene flow between the sampled populations: linearised pairwise Fst was highly correlated with great circle distance (Mantel test $r = 0.79, p = 0.01$; Fig 5.2B).

Whilst the genetic distance between populations was significantly affected by geographic distance, no evidence of an effect of altitude was found in the isolation by resistance analysis (Figure 5.2C). Although it has been proposed that the Rocky Mountains may act as barrier for yellow warblers, increasing genetic structuring [223], changing altitude resistance in CircuitScape did not affect our ability to predict genetic distance between populations. Therefore, I found no evidence to support the impact of topography on population structure in the North American yellow warbler (Figure 5.2C and 5.2D).

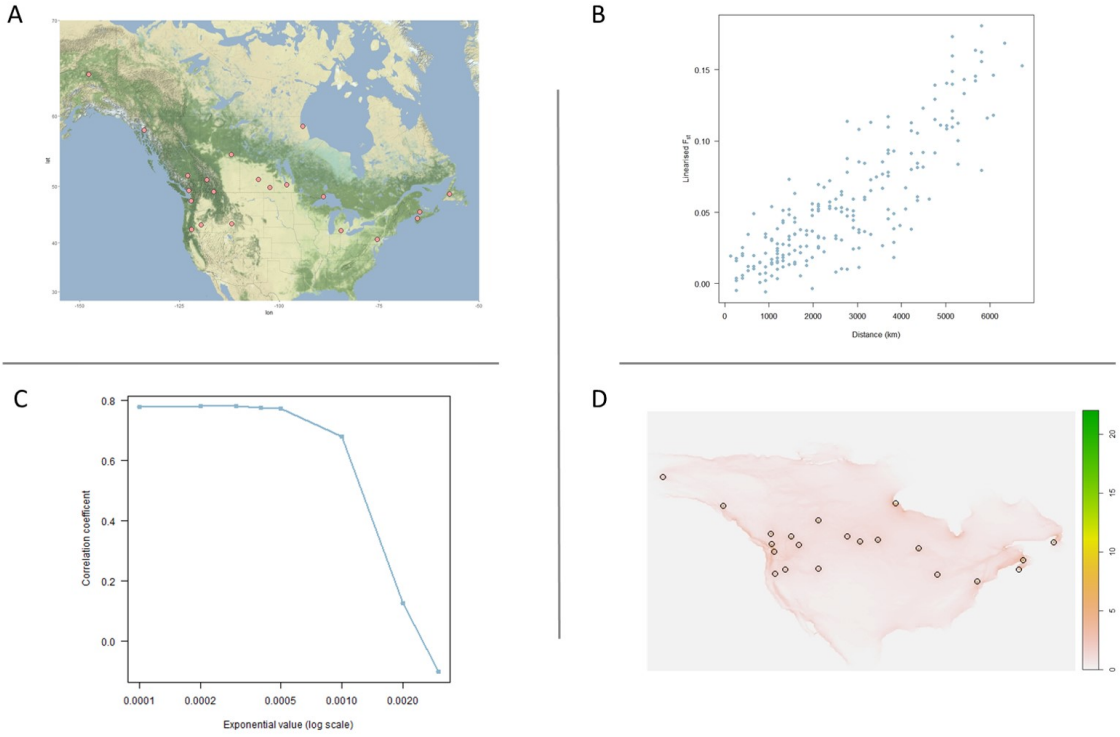


Fig. 5.2 A) Location of all 21 populations sampled. B) Graph of relationship between linearised Fst and distance between each population, the isolation by distance. C) Impact of altering the cost of crossing cells at high altitude, the exponent value (a) alters the effective resistance according to the function $e^{(a \cdot elevation)}$: increasing altitude cost leads to a decrease in correlation between distance and genetic distance. D) CircuitScape connectivity map showing circuit theory 'current flow' between the 21 sampled populations (circled in black).

5.3.2 Fitting CISGeM

Having examined the features and dynamics of the contemporary yellow warbler population, highlighting key patterns any model needs to capture, I fitted a model using the Climatically Informed Spatial Genetic Model (CISGeM) framework. Pairwise plots of the summary statistics distributions from a Monte-Carlo sweep can be found in the supplementary material (D.3). The posterior distributions (Figure 5.3) showed a clear signal the migration rate (m). Within the best fitting models, values of m remained small compared to the values of c. This suggests that the clear isolation by distance pattern seen in the empirical data is best captured by low migration rates after a rapid colonisation of new habitat. Notably, there was little signal from other parameters such as the growth rate of the population (r), however, as one might expect for a small bird, the fitted values of K_0 were very high. Our model indicates effective population sizes of K_0; the single ancestral population used to simulate coalescences beyond the 50,000 generations; of ~2 million (see Table 5.1).

5.3.3 Sequence simulation

MtDNA sequences were generated using parameter values in Table 5.1.

Simulation	c	m	r	*allometric scaling exponent*	*allometric scaling factor*	K_0
77	0.1248	0.0091	0.3427	0.3759	258.32	2137404
71	0.0801	0.0104	0.0391	0.2383	282.02	2433694
20	0.1072	0.0078	0.5577	0.7359	470.45	2027860

Table 5.1 Details of the parameter sets from each of the three simulations chosen for mtDNA generation. Columns are; simulation ID, colonisation rate (c), migration rate (m), growth rate (r), allometric scaling exponent, allometric scaling factor, and ancestral population size estimate K_0.

5.3.4 BSPs

All the BSP profiles recover population sizes that are several orders of magnitude larger than the values from the simulated demography. However, this is to be expected considering that these are not isolated populations and the BSP is therefore likely to be capturing the influence of the larger, global population through migration. To confirm the decline was not being driven by deep structure in the mtDNA genealogy, I investigated the impact of filtering the sequence data with rules similar to those used in previous chapters. Haplotype networks were drawn for one dataset from each of the three populations, generated under Simulation 20 parameter values. Samples that were on branches with > 30 mutations were then dropped

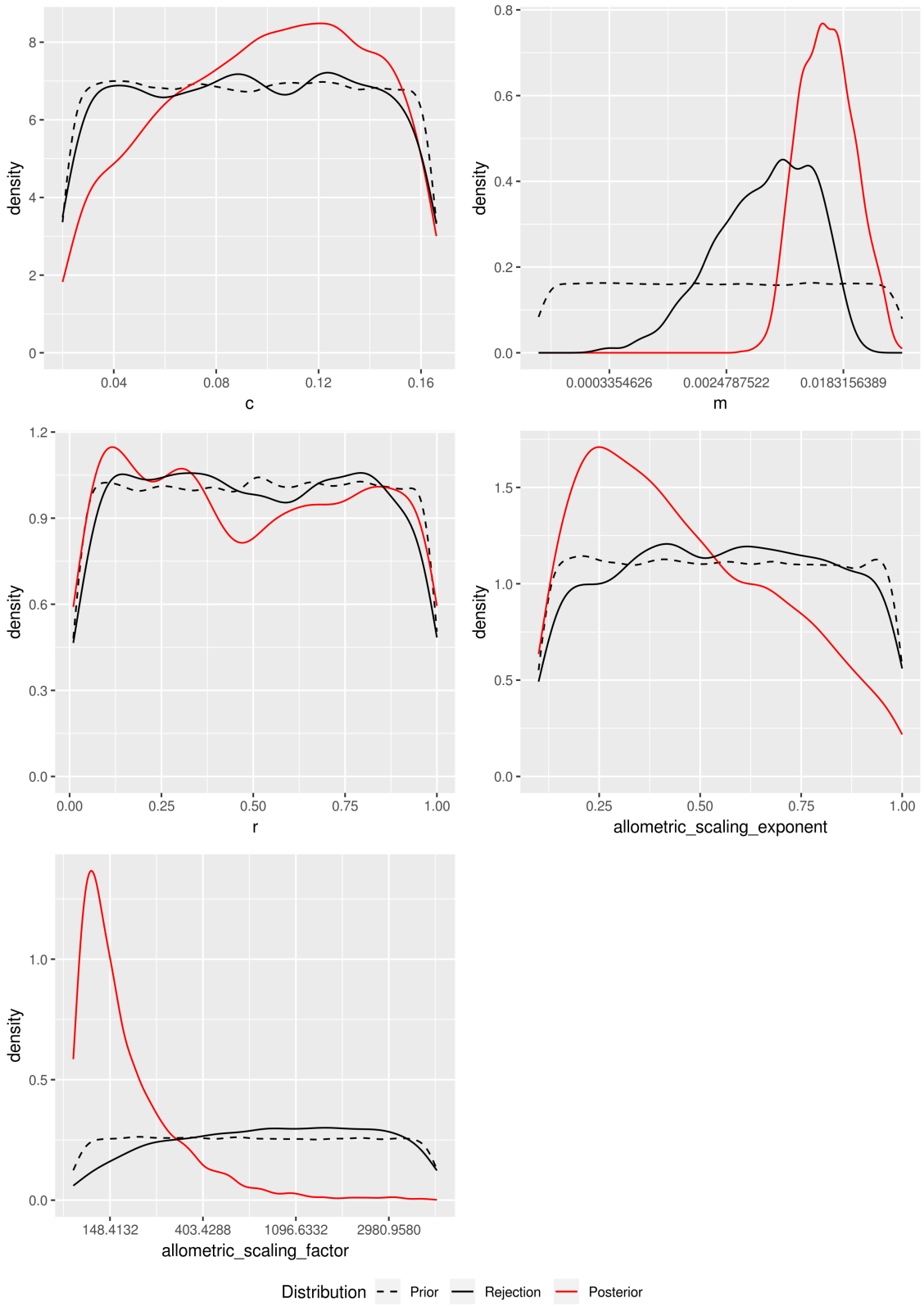


Fig. 5.3 Distribution of key parameters; colonisation rate (c), migration rate (m), growth rate (r), allometric scaling exponent, and allometric scaling factor. Dashed black line shows prior values, solid black line the result of rejection sampling, and the solid red line represents the posterior values from ABC.

(Oregon, $n = 4$; Pennsylvania, $n = 10$; Manitoba, $n = 16$) and these new datasets were run with the same BEAST settings. Though inferred population size was smaller and confidence intervals were tighter (Supplementary Fig. D.6) I found no impact on the overall trends recovered, all further analysis was therefore performed on the full datasets.

Oregon, USA;

With both 1kb of sequence data (Figure 5.4.), representative of one mtDNA gene, and 16kb (Figure 5.5), representative of a full mtDNA genome, BSPs struggle to capture the simulated demographic history for this population. The Oregon population was established 50 kya and has maintained a stable population size since colonisation, yet, both 1kb and 16kb of sequence data recover a decline to the present. Profiles from 1kb of data show a smooth downward trend, whereas the population size change events are more sharply defined with 16kb. Two simulations recover an initial decline around 40-60 kya, and each of the 16kb BSP profiles has a steep decline from ~5 kya. Previous studies have cautioned that recently declining N_e values towards the present maybe artefactual [82] and as genetic data often struggle to capture very recent events I do not consider modern declines to be reliable.

Pennsylvania, USA;

Using a single mtDNA gene, the BSP was unable to capture the bottleneck event this population underwent. With 1kb of data the signal of population decline matches that of the Oregon population profiles, it never captures the depth of decline, nor shows any indication of recovery. Simulation 71 does capture greater population decline than other profiles, though with 1kb of data this initial decline starts very early and appears unrelated to the bottleneck event. However, again, trends are more sharply defined with 16kb of data and here the decline captured by simulation 71 is clearly timed around 50 kya, the start of the simulation.

Manitoba, CA;

This location was colonised by the yellow warbler most recently (~9 kya). For both 1kb and 16kb the BSP population trend was a decline to the present day. For all simulations, when using 1kb of data, this was a smooth decline originating around 25 kya. With 16kb of data the simulations recover more detailed profiles. BSPs from Simulation 71 and 77 indicate initial declines focused around the start of the simulation (50 kya), with further declines around 8-4 kya captured by all the BSPs.

Pooled dataset;

The pooled dataset did not enable the BSP to capture a profile any more detailed than the individual populations and notably this dataset, built of 30 samples from each of the three populations, has much wider CI than any other profile (data not shown). With 16kb of data all the BSPs return a major decline around 40-60 kya. With 1kb, two BSPs (Simulation 20 and Simulation 77) indicate declines around 30-50 kya. The other BSP also indicates a major overall decline in population size, though this decline broken up into two events, one ~150 kya and one ~20 kya. With the 16kb dataset every BSP shows a clear decline 40-50 kya before a, second, very recent decline to the present day. As predicted by other studies, this inherently more structured dataset, seems to have caused the BSP to overestimate the population size further than the single sample site profiles.

5.4 Discussion

To test the effect of expansion dynamics on Bayesian Skyline Plots (BSPs), I produced the first explicit demographic reconstruction of the North American yellow warbler during the last glacial cycle. I then simulated sequence data using realistic parameters and explored how the profiles of changing N_e vary through time depending on the local population history. In the past BSPs have been linked to range changes and assumed to represent or capture the same trends as seen in SDMs. However, previous chapters in this thesis have shown that this assumption is problematic, and local range dynamics may have more of an influence on the genetic history recovered from a population than is commonly supposed.

It was important that a set of realistic parameters were used to generate mtDNA sequences. Generally, the simulations that fitted the empirical data well selected for a rapid rate of colonisation with a comparatively low rate of migration. This combination of values allows the modelled population to quickly expand its range as new habitat becomes available whilst also developing and maintaining a strong IBD, patterns clearly seen in the empirical yellow warbler data. The migration and colonisation features highlighted by this model, therefore, seem to present a credible set of expansions dynamics.

The other strong signal from the model was that of a large population size. This feature make sense in the context of a small, common, passerine bird. When we consider average population sizes for small, common, short lived, passerine species, the large K_0 value, despite being several orders of magnitude bigger than that which was recovered for the global human population, is fitting. I am are therefore satisfied that the model used to generate the BSP input sequence data is adequately capturing the patterns found in the yellow warbler population.

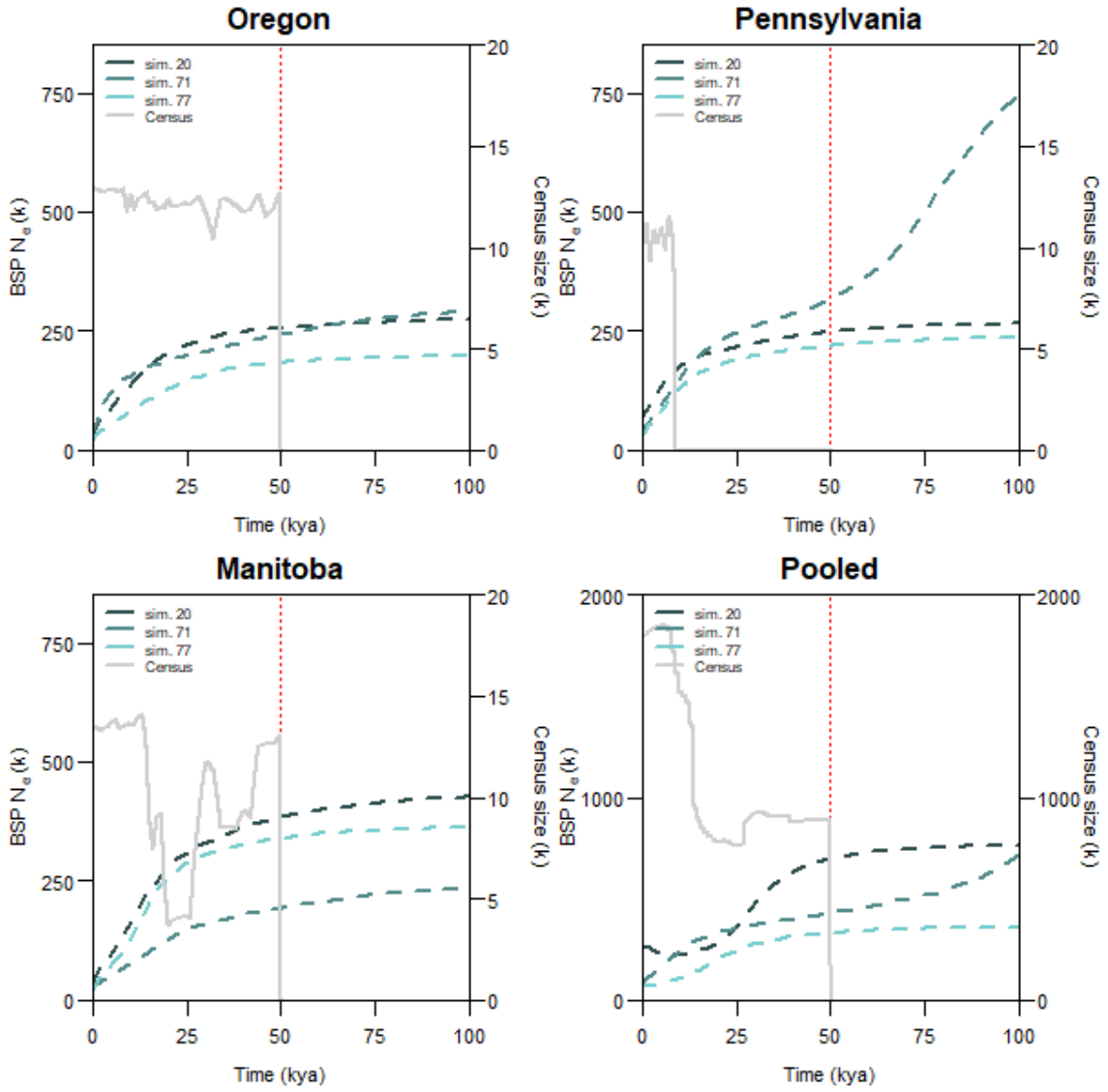


Fig. 5.4 Bayesian Skyline Plots showing the size trend in different yellow warbler populations. The first three BSPs are constructed with samples from a single population ($n = 90$), the fourth BSP was obtained by compiling 30 samples from each of the previous three populations ($n = 90$). The left-hand y axis indicates the effective population size (N_e), the right-hand y axis the simulated census size. For the 'Pooled' population census size is for the full population, drawn from all available demes. The dashed lines are the median estimate for each BSP, the red vertical dotted line indicates the start of the simulation. The x axis is limited to ~100 kya, 50 kya before the simulation began. These profiles were generated from 1kb of mtDNA sequence data.

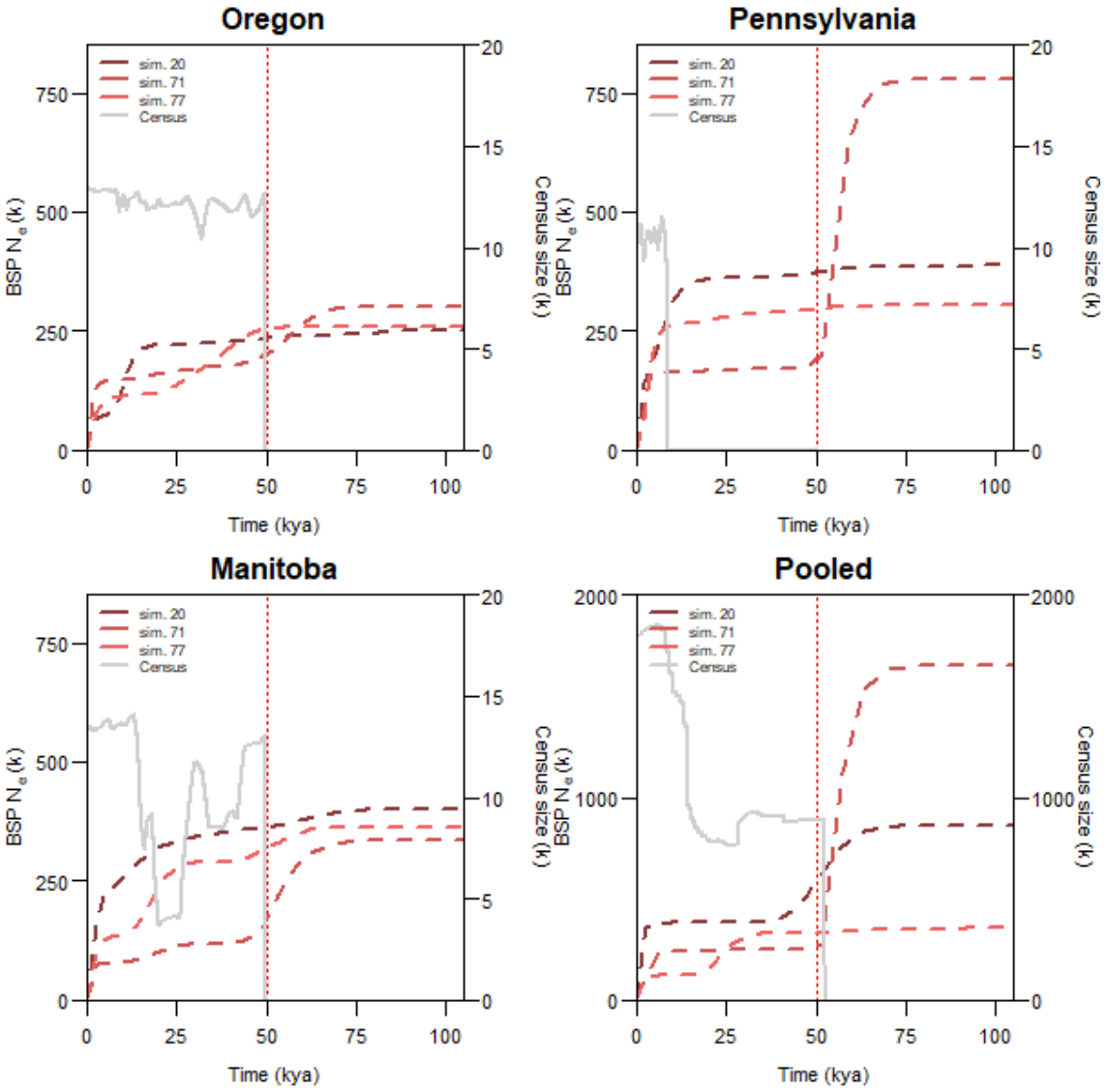


Fig. 5.5 Bayesian Skyline Plots showing the size trend in different yellow warbler populations. The first three BSPs are constructed with samples from a single population ($n = 90$), the fourth BSP was obtained by compiling 30 samples from each of the previous three populations ($n = 90$). The left-hand y axis indicates the effective population size (N_e), the right-hand y axis the simulated census size. For the 'Pooled' population census size is for the full population, drawn from all available demes. The dashed lines are the median estimate for each BSP, the red vertical dotted line indicates the start of the simulation. The x axis is limited to ~100 kya, 50 kya before the simulation began. These profiles were generated from 16kb of mtDNA sequence data.

Whilst I modelled three populations with different, known, demographic histories, and numerous studies have demonstrated that BSPs from different regions of a species habitat return different population histories [174], a large population decline is the dominant signal in all simulated populations. More recent events (e.g. the postglacial colonisation of Manitoba) are not recovered and the unique demographic profiles for each of the three populations are not captured in the BSP. Notably, no BSP for Pennsylvania gives any indication of the major bottleneck that the population experienced ~25 kya, an event that reduced the census N_e by over half. Equally, the timing of major N_e change events for Oregon and Manitoba are not dissimilar, despite the ~40 kya difference in initial expansion time of the modelled census population. Compared to 1kb of sequence data, the use of 16kb of sequence data does appear to sharpen up size changes, making the median N_e for a population change over a shorter period of time.

Despite using the best genetic and environmental data available to me, it is possible that a mismatch in the type of data used to fit the model vs the type of data used to fit the BSP is confounding my analysis. The model is built and fitted to genomic data, whereas I am simulating uniparental mtDNA for BSP analysis. This means that key parameters, such as migration (m), are being drawn from a value averaged across the sexes, yet, features such as sex biased dispersal could mean this parameter fit is not appropriate for the maternally inherited mtDNA. However, in birds, females tend to be the dispersing sex.

I also explored the distribution of coalescent events which revealed that there appears to be two key phases of coalescence. Firstly, a burst of events happen within-deme, then lineages that 'escape' the deme via migration coalesce far later. The spatial model I use works with demes that are discrete spaces, allowing for a step change in coalescent likelihood within the deme and outside the deme. The BSP profiles seem to be being dominated by the events that happen in this first within-deme phase, where a high density of coalescences is interpreted as a significant population bottleneck. One approach that could lessen the strength of this signal could be to move away from the discretisation of space. Indeed, models that incorporate continuous space have recently been developed [224] but there are still major computational challenges to overcome before these tools would be suitable for an area on the scale of this study.

A further aspect of the model that could be hampering the recovery of more accurate BSP profiles surrounds the initialisation of the simulations. Each simulation is seeded from one single deme and so all the lineages must pass through that deme. When exploring the timings of coalescent events, I found that, in a few cases, coalescent events cluster around the start of the simulation, happening within the first few generations before the lineages scatter across the map. Whilst this artefact does not impede the fitting to pairwise diversity from genomic

data, it could confound the features of a BSP. A burst of coalescence events at the beginning of the spatial simulation would be interpreted by skyline methods to indicate a very strong bottleneck event (Figure 5.6), if strong enough this artefactual bottleneck could swamp other demographic signals. Although beyond the scope of my PhD to resolve, one way to tackle this initiation signal could be to begin the simulation with multiple inhabited demes instead of a single deme (e.g. start simulation from Figure 5.1C rather than Figure 5.1A).

Based on this analysis as it stands, it is not possible to draw any conclusions about what population signals BSPs are able to recover, nor what real world signals they may struggle with. If there were a comparable, empirical, mtDNA dataset for the yellow warbler population it might be possible to further diagnose the issues faced in this chapter, but unfortunately these data are not available. In the future it would be interesting to use the approach developed here to investigate a population were both genomic and full mitochondrial data are available. However, for now, there remains an open issue about what is driving the overwhelming pattern of expansion recovered in avian BSPs I identified in Chapter 4.

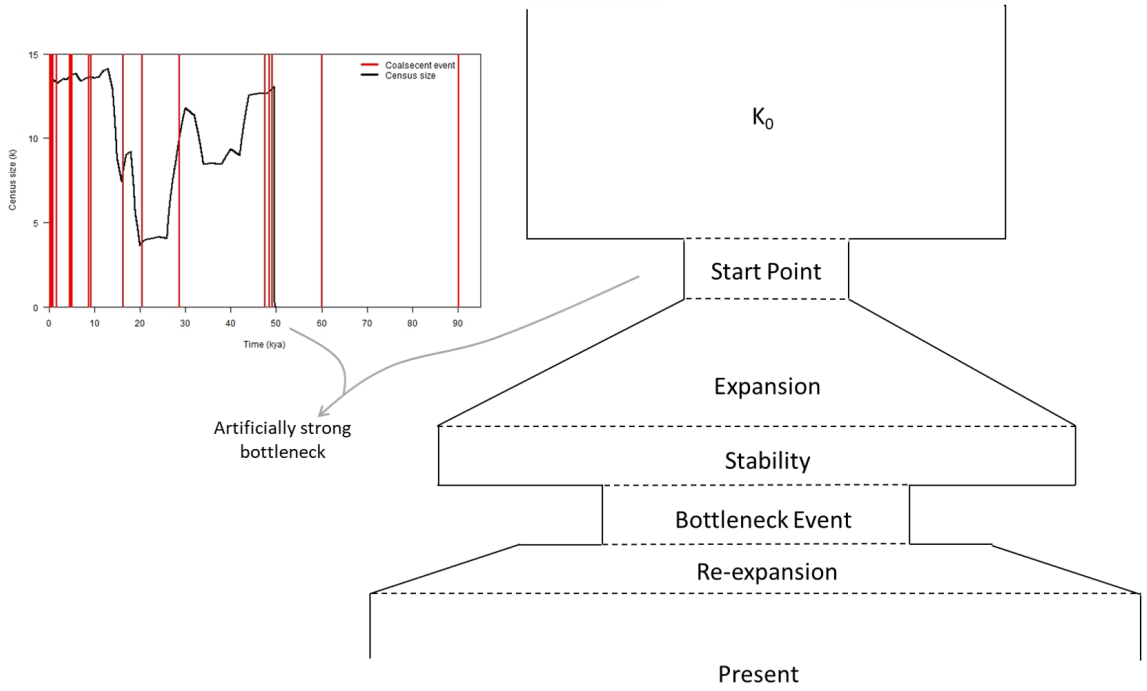


Fig. 5.6 Schematic representation of the issue with a single deme being used to seed the spatial model. As the model reconstructs gene genealogies, moving from the present into the single ancestral population (K_0) in which the distribution of older coalescents are approximated, there will be a clumping of coalescence events. This increased density of events is interpreted as a major bottleneck by BSP analyses. Inset shows an illustrative real case; red lines are coalescent events, black line is the modelled census population history, simulation started at 50 kya.

Acknowledgements

I would like to thank Rachel Bay for help accessing the yellow warbler RAD-seq data.

Chapter 6

General Discussion

6.1 Discussion

During the Pleistocene there were large climatic changes across the world. It has been argued that these climate oscillations had profound impacts on the distribution, demography and diversity of species but getting direct evidence for such impacts is complicated [225]. One solution is to exploit the growing availability of genetic data, coupled with algorithmic developments and increasing computational power, that mean it is now possible to explore the footprint of major climatic events left in the genomes of modern species. Indeed, reconstructing the past dynamics of a population is important for putting modern day species' demographics in context, providing insight into key features such as historic population size, routes of migration and time of speciation (e.g. [226, 227, 228]).

In this thesis I have investigated to what extent we can reconstruct the past using genetic data. I started by showing that even with large numbers of complete mtDNA genomes, the kind of tools that are commonly used to reconstruct population sizes, such as Bayesian skyline plot (BSP) methods, can be difficult to interpret. Then, using publicly accessible sequence data, I explored BSP profiles for ~100 Holarctic bird species, investigating the extent to which BSPs relate to other lines of evidence commonly used to infer past demographic changes. Having identified a potential mismatch in trends between BSPs and Species Distribution Models (SDMs), I wanted to tease apart the factors that could confound or complicate interpretation of coalescent based methods for reconstructing the past. However, the integration of realistic spatial modelling, and the inference of population history from simulated sequence data is complex and further work remains to be done before this proves informative.

6.1.1 Public databases: opportunities, but not without challenges

Today, publicly accessible sequence data offers an extensive resource. Yet, variable quality and a lack of consistency in how data are deposited in databases provide a real challenge in using this resource. At the time of my thesis, there were large amounts of existing public domain mitochondrial data available but only modest numbers of high-quality whole genome sequences for species other than humans. Using existing data significantly diminishes the otherwise high cost of data acquisition and can help to reduce sampling gaps that single studies may face. As the amount of sequence data available in repositories such as GenBank [229] grows, so compiling multi-study datasets is becoming an ever more exciting and

powerful tool. With the cost of genome sequencing continuing to fall and the data policy of many funding bodies requiring that any sequence data generated be released into the public domain, there has been a recent surge in the different types of data becoming available. For example, large data sets of RAD-sequencing and even whole genome sequences for non-model species are becoming more commonplace. Yet, large nuclear DNA (nDNA) datasets are still limited compared to the volume of mitochondrial DNA (mtDNA) data that has been published and so I chose to focus most of my work on mitochondrial data.

Despite the benefits of large database-enabled studies, preparing data that originate from diverse sources is a complex and difficult undertaking, remaining a significant upstream bottleneck for these kind of analyses. Complexities stem from the fact that individual studies start with raw material of a variable standard, use different bioinformatics pipelines, produce sequences of different qualities, explore different parts of the genome (different genes, different regions within the same gene), and different lengths of data. Additionally, open access databases have limited capacity for regulation and quality control. For example, there is no way to confirm the quality of the sequencing work or even to verify the accuracy of associated meta-information such as geolocation data, if it is provided at all. I wanted to investigate a substantial number of species, therefore, I developed a semi-automated analysis pipeline that enables efficient extraction and processing of large amounts of data in order to compile a suitable genetic dataset from which to work.

The power of individual genetic studies can often be restricted due to budget limitations, small sample sizes, poor range coverage or taxonomic gaps. Yet, all these elements can be improved with creative use of existing datasets. As access to both sequencing tools and public sequence repositories grows, so too will the power of analyses that combine data collected for different purposes. I hope that the package and pipeline created during my PhD will prove useful to others aiming to draw on existing sequence data. By providing a methodology, as well as the necessary tools, for the acquisition and, critically, standardisation of candidate molecular datasets from GenBank, my hope is that the mtDNAcomp package can help researchers take full advantage of the extraordinary resource presented by such repositories. Other factors that will be key to increasing the utility of public databases will be regulating the way people report data because, at the moment, the field of populations genetics has no standardised approach for data reporting. By including as much meta-information about a sample as possible within the database, rather than simply alongside a publication, scientist can significantly increase the ease of using that data. Even simple things, like checking for spelling errors and using standardised gene abbreviations, can improve accessibility of the data to other researchers. I hope that, as the community grasps the potential of compiling DNA sequences for tasks secondary to the original reason they were collected, there will be

an increase in care and precision in the data upload process. Simple changes to the quality and consistency of data accessions will benefit a wide range of studies, not only work focused on reconstructing population history.

6.1.2 Reconstructing past demographic changes

One of the key approaches available today for recovering population size change through time, for both extant and extinct species, are skyline plots. Initially I applied the Extended Bayesian Skyline Plot (EBSP) method to a gold standard sequence resource from a model system, human mtDNA from The 1000 Genomes Project Phase 3 data. The combination of a well-studied species and a high-quality dataset allowed me to explore how robust the trends and patterns recovered from each region were, along with how well the recovered trends corroborated features of human population history inferred from other markers and methods.

Whilst I was able to capture enough detail to note a trend for populations from the similar regions to recover similar profiles on a global cline, it was interesting that even with full mtDNA genomes I was unable to recover any signal of the Neolithic, a major event in human history [174]. This highlighted that skyline plots need to be interpreted carefully and with the appropriate temporal and ecological context because major effective population size (Ne) change events, such as bottlenecks or rapid expansions, can erase or swamp other signals, even from complete mtDNA genomes [72]. Moreover, as datasets of this size and quality are uncommon, especially for non-model organisms, problems of signal loss are likely to be exacerbated in other species.

After the human study, I then progressed to analyse data from a non-model species, specifically a range of bird species from the Northern Hemisphere. As sufficient volume of full mitochondrial genomes was not available for many species, I chose to focus on two mtDNA genes frequently sequenced in birds; cytochrome B (cytb) and NADH-dehydrogenase subunit 2 (ND2). These genes, both approximately ~1000 base pairs, can obviously hold fewer mutations and thus less information than a full genome. However, as the mtDNA genome is non-recombining it must be treated as a single locus and so, whilst addition sequence length is helpful for resolving population histories, the use of only one mtDNA gene instead of a full genome does not result in the same loss of data as excluding multiple other loci from the analysis.

Although, as previously mentioned, combining data from multiple studies can help to improve features of individual datasets, real datasets will still invariably be susceptible to problems including variable sequence quality and population structure. In addition, data collated from the public domain can be confounded by factors such as sample mislabelling, poorly documented records, and unstructured sampling strategy. As a result, when handling

many datasets built from data collected for diverse reasons other than investigating population history, MCMC convergence can be problematic. In order to process >160 datasets within one comparable, like for like pipeline it was necessary for me to sacrifice ultimate power on an individual level for robustness and so I moved away from the more flexible EBSP, going back to the BSP approach. This allowed me to control and define more parameters, better constraining the runs and helping to achieve convergence.

6.1.3 Reconstructing the past is difficult

By exploring so many profiles together my work has allowed a broader view in which many datasets can be visualised side by side. When this is done, common problems become much more apparent. For example, 98% of the avian profiles I constructed showed an overall increase in population size, despite species having diverse ecologies and contrasting range dynamics. If we look at the spread of profiles that we get within any one broad habitat class these include a range in which there may be a trend for, for example, an increase to occur at a given time, but with much variation about this. Individual profiles within this range may appear very convincing in isolation, but when grouped like this it becomes clear that a naïve interpretation is inappropriate, as there are obviously additional factors confounding and complicating the analysis.

Previous studies have flagged that skyline plots require careful interpretation, yet, problematic datasets and flawed interpretations remain commonplace in published work with storytelling appearing to be a critical driver in some studies. For example, if there are insufficient mutations in the input data to reconstruct an informative history, a flat, or overly simplified, signal may be returned. Scenarios that reduce the information within a dataset, such as bottleneck events that decrease overall diversity, poor sampling strategy that fails to cover a populations' range, or sequencing of insufficient base pairs given the rate of mutation, are not uncommon. However, flat profiles with wide confidence intervals are frequently presented as being indicative of stable population size. Whilst this may be genuine, it is rarely acknowledged that this shape of profile could also be recovered by a number of other scenarios, or may be artefactual [72, 230].

Even if skyline profiles are capturing a true demographic history they must be set in an appropriate geographical and temporal context for interpretation. Looking at both the human populations in Chapter 2 and the demographic history of the yellow warbler in Chapter 5, it was noticeable that the recovered trends do not necessarily capture the history of the sampled site itself. When looking at these profiles, single, dominating events could be a source of disconnect between a reconstructed demography and paleoenvironmetal history, confounding our understanding of the factors that drove population change. If major events

swamp other signals, a user can never be sure that the absence of a feature means the feature was not present or simply that the signal of that feature has been lost. This issue will not be specific to skyline methods, and other approaches such as sequence mismatch analysis (MMA) and pairwise sequentially Markovian coalescent (PSMC) are likely to suffer from the same limitation. Despite this, the wider environmental and historical setting of the sampled population is rarely discussed when interpreting population profiles.

It is important to note that the difficulties I had with capturing the population dynamics of the yellow warbler may, to some degree, be due to artefacts associated with my spatial modelling and sequence generation. Factors such as a high density of coalescent events occurring at the very start of the simulation, or an unusually high level of structuring in the generated sequences, are likely to have confounded the BSP analysis giving the impression of major bottleneck events and / or disproportionately large historical populations. Although beyond the scope of this PhD, future work could ease some of these methodological challenges making it possible to revisit this approach to exploring the difficulties that face BSPs and BSP interpretation. For example, changing the way the simulation framework initially seeds the model would be a great step towards building an even more realistic temporal demographic model.

The dangers of overinterpretation also lurk for studies that attempt to draw broad conclusions extrapolated from one species: this could be problematic because any given single profile has the potential to be noisy, and thus misleading. This is best demonstrated with the avian profiles I investigated. Significant trends in the timing of expansion after the LGM exist between the broad habitat groups considered, yet, some individual species have a time of increase that is very different from other species from the same habitat. Although it remains unclear the extent to which these reflect genuine biological signals or spurious effects such as incorrect species-specific mutation rates, it is easy to see that erroneous or incomplete interpretations about the ecosystem or community could be drawn. Therefore, it is potentially concerning that a search of Web of Science using the key words ‘Bayesian Skyline Plot’ revealed that, bar two, all of the studies applying BSP analysis published so far in 2019 ($n = 26$) are based on data from a single species.

6.1.4 Future direction

In summary, exciting tools are available to make use of the wealth of genetic data available to us now and in the coming decades. However, we still need to better understand how real population dynamics can influence the profiles recovered and what this means for our interpretation of these analyses. It is critical that the rapidly evolving methods and tools being developed are not applied blindly under the assumption that ‘genetics never lie’. My

work, specifically in Chapter 2 and Chapter 4, shows that there is real value in using multiple lines of evidence. It has reiterated the importance of considering that demographic trends recovered from one method may not tell the full, or only, story about a population's history. That alternative narratives and additional support are presented and discussed is critical, whatever the approach used, because all models have assumptions and it is often hard to balance these assumptions appropriately. For example, within this thesis, Chapter 5 shows that a complex demography might lead to patterns that are very difficult to interpret. Yet, without the prior knowledge of the population history that I had in the yellow warbler dataset (since I simulated it), it is possible that a naïve user investigating this single species would have assumed that true population patterns had been captured and a narrative could have been built around these recovered skyline profiles.

As a field, genetics still faces a number of challenges that need to be overcome before unhelpful habits become further ingrained and bad practise normalised. Whilst, compared to other fields, genetics is reasonably far forward in mandating the availability and sharing of datasets, the way data is stored is often inconsistent. Although sequencing data are, to some degree, intrinsically noisy [231] and, in reality, quality will always be down to the user, the current level of freedom surrounding data uploads allows for the introduction of an excessive amount of human error and idiosyncrasies. Small changes to major repositories that would not significantly hamper the process of making data available, such as the use of radio-buttons or drop-down menus rather than open text fields, would have a huge impact on the usability of publicly accessible sequence data. The exciting opportunity to build unprecedented datasets from existing data is only feasible programmatically. Yet, it is not trivial to scrape information that is not in a standard format, could include human errors such as spelling and formatting issues, and may have missing fields. To ensure published data are truly discoverable for others to use, and to safeguard basic standards of reproducible science, a greater level of standardisation urgently needs to be implemented.

Further, issues surround the accessibility and standardisation of metadata. Sequence data can be highly informative, yet, accurate, complete, and intelligible supporting information about features such as sample sites, sample sizes, and sequencing methods, is vital to make this data usable for future studies. Depending on the question being investigated, different aspects of metadata may become pertinent. If users only make available half of the information about a sample or dataset there is, not only a huge loss of information but also a potential restriction in the utility of that data for further studies. Focusing on only two genes from avian mtDNA uploaded to one database, I found a huge divergence in the approach authors took to metadata, and indeed, a lack of supporting information was a common reason for data exclusion.

As the era of whole genome data dawns, I feel that the importance of standardisation and regulation is only going to grow. Whilst moving to whole genome data will ease some issues around compiling large multi-study dataset, these data have the potential to be equally, if not more, inconsistent. On one hand, if raw reads are available, it is possible to start an analysis afresh, realigning the data and controlling the processing yourself. But this is a big and expensive job that is ideally avoided. However, data that is already processed may have undergone a whole host of different, possibly incomparable, bioinformatics processes. If the bioinformatics tools and processes used to create and format the data are not reported alongside the data themselves, it is, at best, a significant inconvenience to a secondary user, but, at worst, there is a serious risk of the permanent loss of extensive information. Whilst certain aspects, such as file formats, are standardised, most parts of the multistep bioinformatics pipeline remain undirected and cannot be assumed. Tackling this issue won't be easy but, if the field of genetics wants to extract full value from the huge volume of sequence data now available, it is critical that solutions are implemented before the volume of mismatched or underreported data becomes insurmountable.

References

[1] Paul J. Crutzen. “Geology of mankind: the Anthropocene”. In: *Nature* 415 (2002), p. 23.

[2] Will Steffen, J Crutzen, and John R. McNeill. “The Anthropocene: are humans now overwhelming the great forces of Nature?” In: *Ambio* 36.8 (Dec. 2007), pp. 614–21. DOI: 10.1579/0044-7447(2007)36[614:taahno]2.0.co;2.

[3] Simon L. Lewis and Mark A. Maslin. “Defining the Anthropocene”. In: *Nature* 519.7542 (2015), pp. 171–180. DOI: 10.1038/nature14258.

[4] Rodolfo Dirzo, Hillary S. Young, Mauro Galetti, Gerardo Ceballos, Nick J. B. Isaac, and Ben Collen. “Defaunation in the Anthropocene”. In: *Science* 345.6195 (July 2014), pp. 401–406. DOI: 10.1126/science.1251817.

[5] Anthony D. Barnosky, Nicholas Matzke, Susumu Tomiya, et al. “Has the Earth’s sixth mass extinction already arrived?” In: *Nature* 471.7336 (Mar. 2011), pp. 51–57. DOI: 10.1038/nature09678.

[6] Stuart L Pimm, Gareth J Russell, John L Gittleman, and Thomas M Brooks. “The Future of Biodiversity”. In: *Science* 269.5222 (July 1995), pp. 347–350. DOI: 10.1126/science.269.5222.347.

[7] Stuart L. Pimm, Clinton N. Jenkins, Robin Abell, et al. “The biodiversity of species and their rates of extinction, distribution, and protection”. In: *Science* 344.6187 (May 2014), pp. 1246752–1246752. DOI: 10.1126/science.1246752.

[8] Godfrey Hewitt. “The genetic legacy of the Quaternary ice ages”. In: *Nature* 405.6789 (June 2000), pp. 907–913. DOI: 10.1038/35016000.

[9] Thomas B. Chalk, Mathis P. Hain, Gavin L. Foster, et al. “Causes of ice age intensification across the Mid-Pleistocene Transition”. In: *Proceedings of the National Academy of Sciences* 114.50 (Dec. 2017), pp. 13114–13119. DOI: 10.1073/pnas.1702143114.

[10] A. Rus Hoelzel. “Looking backwards to look forwards: conservation genetics in a changing world”. In: *Conservation Genetics* 11.2 (2010), pp. 655–660. DOI: 10.1007/s10592-010-0045-4.

[11] Sune O. Rasmussen, Matthias Bigler, Simon P. Blockley, et al. “A stratigraphic framework for abrupt climatic changes during the Last Glacial period based on three synchronized Greenland ice-core records: Refining and extending the INTIMATE event stratigraphy”. In: *Quaternary Science Reviews* 106 (2014), pp. 14–28. DOI: 10.1016/j.quascirev.2014.09.007.

[12] Greenland Ice-core Project (GRIP) Members. “Climate instability during the last interglacial period recorded in the GRIP ice core”. In: *Nature* 364.6434 (July 1993), pp. 203–207. DOI: 10.1038/364203a0.

[13] Willi Dansgaard, Sigfús J. Johnsen, Henrik B. Clausen, et al. “Evidence for general instability of past climate from a 250-kyr ice-core record”. In: *Nature* 364.6434 (July 1993), pp. 218–220. DOI: 10.1038/364218a0.

[14] Jean Jouzel, Claude Lorius, Jean-Robert Petit, Christophe Genthon, Nartsiss I. Barkov, Vladamir M. Kotlyakov, and Vladimir M. Petrov. “Vostok ice core: a continuous isotope temperature record over the last climatic cycle (160,000 years)”. In: *Nature* 329.6138 (Oct. 1987), pp. 403–408. DOI: 10.1038/329403a0.

[15] Fabrice Lambert, Barbara Delmonte, Jean-Robert Petit, et al. “Dust-climate couplings over the past 800,000 years from the EPICA Dome C ice core”. In: *Nature* 452.7187 (Apr. 2008), pp. 616–619. DOI: 10.1038/nature06763.

[16] Jean Jouzel, Valerie Masson-Delmotte, Olivier Cattani, et al. “Orbital and Millennial Antarctic Climate Variability over the Past 800,000 Years”. In: *Science* 317.5839 (Aug. 2007), pp. 793–796. DOI: 10.1126/science.1141038.

[17] Fraser J G Mitchell. “Exploring vegetation in the fourth dimension”. In: *Trends in Ecology and Evolution* 26.1 (2011), pp. 45–52. DOI: 10.1016/j.tree.2010.10.007.

[18] Pavel E. Tarasov, Elena V. Bezrukova, and Sergey K. Krivonogov. “Late glacial and holocene changes in vegetation cover and climate in southern Siberia derived from a 15 kyr long pollen record from Lake Kotokel”. In: *Climate of the Past* 5.3 (2009), pp. 285–295. DOI: 10.5194/cp-5-285-2009.

[19] Heikki Seppä. *Pollen Analysis, Principles*. 2nd ed. August. Elsevier B.V., 2013, pp. 794–804. DOI: 10.1016/B978-0-444-53643-3.00171-0.

[20] Thomas Giesecke, Basil Davis, Simon Brewer, et al. “Towards mapping the late Quaternary vegetation change of Europe”. In: *Vegetation History and Archaeobotany* 23.1 (2014), pp. 75–86. DOI: 10.1007/s00334-012-0390-y.

[21] Thomas Giesecke, Simon Brewer, Walter Finsinger, Michelle Leydet, and Richard H.W. Bradshaw. “Patterns and dynamics of European vegetation change over the last 15,000 years”. In: *Journal of Biogeography* 44.7 (2017), pp. 1441–1456. DOI: 10.1111/jbi.12974.

[22] Jason L. Brown, Daniel J. Hill, Aisling M. Dolan, Ana C. Carnaval, and Alan M. Haywood. “Paleoclim, high spatial resolution paleoclimate surfaces for global land areas”. In: *Scientific Data* 5 (2018), pp. 1–9. DOI: 10.1038/sdata.2018.254.

[23] Matheus Souza Lima-Ribeiro. “EcoClimate: a database of climate data from multiple models for past, present, and future for macroecologists and biogeographers”. In: *Biodiversity Informatics* 10 (2015), pp. 1–21. DOI: 10.17161/bi.v10i0.4955.

[24] Robert J. Hijmans, Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis. “Very high resolution interpolated climate surfaces for global land areas”. In: *International Journal of Climatology* 25.15 (2005), pp. 1965–1978. DOI: 10.1002/joc.1276.

[25] Robert M Beyer, Mario Krapp, and Andrea Manica. “High-resolution terrestrial climate , bioclimate and vegetation for the last 120,000 years”. In: *EarthArXiv* (2019).

[26] Robert S. Sommer and N. Benecke. “The recolonization of Europe by brown bears Ursus arctos Linnaeus, 1758 after the Last Glacial Maximum”. In: *Mammal Review* 35.2 (2005), pp. 156–164. DOI: 10.1111/j.1365-2907.2005.00063.x.

[27] Robert S. Sommer and Frank E. Zachos. “Fossil evidence and phylogeography of temperate species: ’Glacial refugia’ and post-glacial recolonization”. In: *Journal of Biogeography* 36.11 (2009), pp. 2013–2020. DOI: 10.1111/j.1365-2699.2009.02187.x.

[28] Godfrey M. Hewitt. “Genetic consequences of climatic oscillations in the Quaternary”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 359.1442 (2004), pp. 183–195. DOI: 10.1098/rstb.2003.1388.

[29] Godefrey M. Hewitt. “Post-glacial re-colonization of European biota”. In: *Biological Journal of the Linnean Society* 68.1-2 (Sept. 1999), pp. 87–112. DOI: 10.1111/j.1095-8312.1999.tb01160.x.

[30] Antoine Guisan and Wilfried Thuiller. “Predicting species distribution: Offering more than simple habitat models”. In: *Ecology Letters* 8.9 (2005), pp. 993–1009. DOI: 10.1111/j.1461-0248.2005.00792.x.

[31] Jane Elith, Robert P Anderson, Miroslav Dudík, et al. “Novel methods improve prediction of species’ distributions from occurrence data”. In: *Ecography* 29.2 (2006), pp. 129–151. DOI: 10.1111/j.2006.0906-7590.04596.x.

[32] Robert P. Anderson. “Real vs . Artefactual Absences in Species Distributions : Tests for Oryzomys albigularis (Rodentia : Muridae) in Venezuela”. In: *Journal of Biogeography* 30.4 (2016), pp. 591–605. DOI: 10.1039/c0ee00071j.

[33] Alexandre H. Hirzel, Jacques Hausser, Daniel Chessel, and Nicolas Perrin. “Ecological-Niche Factor Analysis : How to Compute Habitat-Suitability Maps without Absence Data?” In: *Ecology* 83.7 (2002), pp. 2027–2036.

[34] Trevor Hastie and Robert Tibshirani. “Exploring the Nature of Covariate Effects in the Proportional Hazards Model”. In: *Biometrics* 46.4 (Dec. 1990), p. 1005. DOI: 10.2307/2532444.

[35] Trevor Hastie and Robert Tibshirani. “Generalized Additive Models”. In: *Statistical Science* 1.3 (Aug. 1986), pp. 297–310. DOI: 10.1214/ss/1177013604.

[36] Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. “Maximum entropy modeling of species geographic distributions”. In: *Ecological Modelling* 190.3-4 (2006), pp. 231–259. DOI: 10.1016/j.ecolmodel.2005.03.026.

[37] Antoine Guisan, Wilfried Thuiller, and Niklaus E. Zimmermann. *Habitat Suitability and Distribution Models*. Cambridge: Cambridge University Press, 2017. DOI: 10.1017/9781139028271.

[38] Miguel B. Araújo and Mark New. “Ensemble forecasting of species distributions”. In: *Trends in Ecology and Evolution* 22.1 (2007), pp. 42–47. DOI: 10.1016/j.tree.2006.09.010.

[39] Andrew Townsend Peterson. “Ecological niche conservatism: A time-structured review of evidence”. In: *Journal of Biogeography* 38.5 (2011), pp. 817–827. DOI: 10.1111/j.1365-2699.2010.02456.x.

[40] Andrew Townsend Peterson, Jorge Soberón, and Victor Sánchez-Cordero. “Conservatism of ecological niches in evolutionary time”. In: *Science* 285.5431 (1999), pp. 1265–1267. DOI: 10.1126/science.285.5431.1265.

[41] Miguel B. Araújo and Richard G. Pearson. "Equilibrium of species' distributions with climate". In: *Ecography* 28.5 (Oct. 2005), pp. 693–695. DOI: 10.1111/j.2005.0906-7590.04253.x.

[42] William B. Monahan. "A mechanistic niche model for measuring species' distributional responses to seasonal temperature gradients". In: *PLoS ONE* 4.11 (2009). DOI: 10.1371/journal.pone.0007921.

[43] Wilfried Thuiller, Laura J. Pollock, Maya Gueguen, and Tamara Münkemüller. "From species distributions to meta-communities". In: *Ecology Letters* 18.12 (Dec. 2015). Ed. by Howard Cornell, pp. 1321–1328. DOI: 10.1111/ele.12526.

[44] Damaris Zurell, Laura J. Pollock, and Wilfried Thuiller. "Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments?" In: *Ecography* 41.11 (Nov. 2018), pp. 1812–1819. DOI: 10.1111/ecog.03315.

[45] Yun-Xin Fu. "A phylogenetic estimator of effective population size or mutation rate". In: *Genetics* 136.2 (1994), pp. 685–692.

[46] Sean Nee, Eddie C Holmes, Andrew Rambaut, and Paul H Harvey. "Inferring Population History from Molecular Phylogenies". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 349.1327 (1995), pp. 25–31.

[47] John F. C. Kingman. "The coalescent". In: *Stochastic Processes and their Applications* 13 (1982), pp. 235–248.

[48] Fumio Tajima. "Evolutionary relationship of DNA sequeences in finite populations". In: *Genetics* 105.2 (1983), pp. 437–460.

[49] Richard R. Hudson. "Testing the Constant-Rate Neutral Allele Model with Protein Sequence Data". In: *Evolution* 37.1 (1983), pp. 203–217.

[50] Oliver G. Pybus, Edward C. Holmes, and Paul H. Harvey. "The mid-depth method and HIV-1: A practical approach for testing hypotheses of viral epidemic history". In: *Molecular Biology and Evolution* 16.7 (1999), pp. 953–959. DOI: 10.1093/oxfordjournals.molbev.a026184.

[51] Oliver G Pybus, Andrew Rambaut, and Paul H Harvey. "An integrated framework for the inference of viral population history from reconstructed genealogies." In: *Genetics* 155.3 (July 2000), pp. 1429–37.

[52] Stephan Schiffels and Richard Durbin. "Inferring human population size and separation history from multiple genome sequences". In: *Nature Publishing Group* 46.8 (2014), pp. 919–925. DOI: 10.1038/ng.3015.

[53] John F. C. Kingman. "On the Genealogy of Large Populations". In: *Journal of Applied Probability* 19.1982 (1982), p. 27. DOI: 10.2307/3213548.

[54] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, 2004.

[55] Richard C. Griffiths and Simon Tavaré. "Sampling theory for neutral alleles in a varying environment". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* (1994), pp. 403–410.

[56] Peter Donnelly and Simon Tavaré. “Coalescents and Genealogical Structure under Neutrality”. In: *Annual Review of Genetics* 29.1 (Jan. 1995), pp. 401–421. DOI: 10.1146/annurev.genet.29.1.401.

[57] William Stewart Grant. “Problems and cautions with sequence mismatch analysis and Bayesian skyline plots to infer historical demography”. In: *Journal of Heredity* 106.4 (2015), pp. 333–346. DOI: 10.1093/jhered/esv020.

[58] Alan R Rogers and Henry Harpending. “Population growth makes waves in the distribution of pairwise genetic differences.” In: *Molecular Biology and Evolution* 9.July (1992), pp. 552–569. DOI: 10.1093/oxfordjournals.molbev.a040727.

[59] Montgomery Slatkin and Richard R. Hudson. “Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations”. In: *Genetics* 129.2 (1991), pp. 555–562.

[60] Alexei. J. Drummond, Andrew Rambaut, Beth Shapiro, and Oliver G. Pybus. “Bayesian coalescent inference of past population dynamics from molecular sequences”. In: *Molecular Biology and Evolution* 22.5 (2005), pp. 1185–1192. DOI: 10.1093/molbev/msi103.

[61] Alan R. Rogers. “Genetic Evidence for a Pleistocene Population Explosion”. In: *Evolution* 49.4 (Aug. 1995), p. 608. DOI: 10.2307/2410314.

[62] Laurent Excoffier and Heidi E L Lischer. “Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows”. In: *Molecular Ecology Resources* 10.3 (2010), pp. 564–567. DOI: 10.1111/j.1755-0998.2010.02847.x.

[63] Vladimir N. Minin, Erik W. Bloomquist, and Marc A. Suchard. “Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics”. In: *Molecular Biology and Evolution* 25.7 (2008), pp. 1459–1471. DOI: 10.1093/molbev/msn090.

[64] Tanja Stadler, D. Kuhnert, Sebastian Bonhoeffer, and Alexei J Drummond. “Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)”. In: *Proceedings of the National Academy of Sciences* 110.1 (Jan. 2013), pp. 228–233. DOI: 10.1073/pnas.1207965110.

[65] Rainer Opgen-Rhein, Ludwig Fahrmeir, and Korbinian Strimmer. “Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo”. In: *BMC Evolutionary Biology* 5 (2005), pp. 1–13. DOI: 10.1186/1471-2148-5-6.

[66] Joseph Heled and Alexei J Drummond. “Bayesian inference of population size history from multiple loci”. In: *BMC Evolutionary Biology* 15 (2008), pp. 1–15. DOI: 10.1186/1471-2148-8-289.

[67] Gustavo Sanchez, Satoshi Tomano, Carmen Yamashiro, Ricardo Fujita, Toshie Wakabayashi, Mitsuo Sakai, and Tetsuya Umino. “Population genetics of the jumbo squid Dosidicus gigas (Cephalopoda: Ommastrephidae) in the northern Humboldt Current system based on mitochondrial and microsatellite DNA markers”. In: *Fisheries Research* 175 (2016), pp. 1–9. DOI: 10.1016/j.fishres.2015.11.005.

[68] Mathias Stiller, Gennady Baryshnikov, Hervé Bocherens, et al. "Withering away-25,000 years of genetic decline preceded cave bear extinction". In: *Molecular Biology and Evolution* 27.5 (2010), pp. 975–978. DOI: 10.1093/molbev/msq083.

[69] Irene Villalta, Fernando Amor, Juan A. Galarza, et al. "Origin and distribution of desert ants across the Gibraltar Straits". In: *Molecular Phylogenetics and Evolution* 118.September 2017 (2018), pp. 122–134. DOI: 10.1016/j.ympev.2017.09.026.

[70] Takahiro Segawa, Nozomu Takeuchi, Koji Fujita, Vladimir B. Aizen, Eske Willerslev, and Takahiro Yonezawa. "Demographic analysis of cyanobacteria based on the mutation rates estimated from an ancient ice core". In: *Heredity* (2018), pp. 1–12. DOI: 10.1038/s41437-017-0040-3.

[71] Simon Y W Ho and Beth Shapiro. "Skyline-plot methods for estimating demographic history from nucleotide sequences". In: *Molecular Ecology Resources* 11.3 (2011), pp. 423–434. DOI: 10.1111/j.1755-0998.2011.02988.x.

[72] W. Stewart Grant, Ming Liu, TianXiang Gao, and Takashi Yanagimoto. "Limits of Bayesian skyline plot analysis of mtDNA sequences to infer historical demographies in Pacific herring (and other species)". In: *Molecular Phylogenetics and Evolution* 65.1 (Oct. 2012), pp. 203–212. DOI: 10.1016/j.ympev.2012.06.006.

[73] Gilean A.T. McVean and Niall J. Cardin. "Approximating the coalescent with recombination". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459 (2005), pp. 1387–1393. DOI: 10.1098/rstb.2005.1673.

[74] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357 (2011), pp. 493–496. DOI: 10.1038/nature10231.

[75] Jonathan Terhorst, John A Kamm, and Yun S Song. "Robust and scalable inference of population history from hundreds of unphased whole genomes". In: *Nature Genetics* 49.2 (Feb. 2017), pp. 303–309. DOI: 10.1038/ng.3748.

[76] Olivier Mazet, Willy Rodríguez, Simona Grusea, Simon Boitard, and Lounès Chikhi. "On the importance of being structured: instantaneous coalescence rates and a re-evaluation of human evolution". In: *Heredity* (Nov. 2016). Ed. by Intergovernmental Panel on Climate Change, pp. 362–371. DOI: 10.1017/CBO9781107415324.004.

[77] Donald Rubin. "Bayesianly justifiable and relevant frequency calculations for the applied statistician". In: *The Annals of Statistics* 12.4 (1984), pp. 1151–1172.

[78] Simon Tavaré, David J. Balding, R. C. Griffiths, and Peter Donnelly. "Inferring coalescence times from DNA sequence data". In: *Genetics* 145.2 (1997), pp. 505–518. DOI: 10.1126/sciadv.1501177.

[79] Jonathan K. Pritchard, Mark T. Seielstad, Anna Perez-Lezaun, and Marcus W. Feldman. "Population growth of human Y chromosomes: A study of y chromosome microsatellites". In: *Molecular Biology and Evolution* 16.12 (1999), pp. 1791–1798. DOI: 10.1093/oxfordjournals.molbev.a026091.

[80] Jean Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. "Approximate Bayesian computational methods". In: *Statistics and Computing* 22.6 (2012), pp. 1167–1180. DOI: 10.1007/s11222-011-9288-2.

[81] Olivier Mazet, Willy Rodríguez, and Lounès Chikhi. "Demographic inference using genetic data from a single individual: Separating population size variation from population structure". In: *Theoretical Population Biology* 104 (2015), pp. 46–58. DOI: 10.1016/j.tpb.2015.06.003.

[82] Rasmus Heller, Lounes Chikhi, and Hans Redlef Siegismund. "The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History". In: *PLoS ONE* 8.5 (May 2013). Ed. by Thomas Mailund, e62992. DOI: 10.1371/journal.pone.0062992.

[83] Wesley M Brown, Matthew George, and Allan C Wilson. "Rapid evolution of animal mitochondrial DNA." In: *Annual Review of Ecology and Systematics* 18.1 (1979), pp. 269–292. DOI: 10.1146/annurev.es.18.110187.001413.

[84] George S. Michaels, William W. Hauswirth, and Philip J. Laipis. "Mitochondrial DNA Copy Number in Bovine Oocytes and Somatic". In: *Developmental Biology* 251.94 (1982), pp. 246–251.

[85] Carmela Gissi, Francesco Iannelli, and Graziano Pesole. "Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species". In: *Heredity* 101.4 (Oct. 2008), pp. 301–320. DOI: 10.1038/hdy.2008.62.

[86] John C. Avise, Jonathan Arnold, R. Martin Ball, Eldredge Bermingham, Trip Lamb, Joseph E. Neigel, Carol A. Reeb, and Nancy C. Saunders. "Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics". In: *Annual Review of Ecology and Systematics* 18.1 (Nov. 1987), pp. 489–522. DOI: 10.1146/annurev.es.18.110187.002421.

[87] Hervé Philippe, Henner Brinkmann, Dennis V. Lavrov, D. Timothy J. Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. "Resolving difficult phylogenetic questions: Why more sequences are not enough". In: *PLoS Biology* 9.3 (2011). DOI: 10.1371/journal.pbio.1000602.

[88] James H. Degnan and Noah A. Rosenberg. "Gene tree discordance, phylogenetic inference and the multispecies coalescent". In: *Trends in Ecology and Evolution* 24.6 (2009), pp. 332–340. DOI: 10.1016/j.tree.2009.01.009.

[89] Erich D Jarvis, Siavash Mirarab, Andre. J. Aberer, et al. "Whole-genome analyses resolve early branches in the tree of life of modern birds". In: *Science* 346.6215 (Dec. 2014), pp. 1320–1331. DOI: 10.1126/science.1253451.

[90] Guojie Zhang. "Genomics: Bird sequencing project takes off". In: *Nature* 522.7554 (2015), p. 34. DOI: 10.1038/522034d.

[91] Ryan C. Garrick, Isabel A.S. Bonatelli, Chaz Hyseni, et al. "The evolution of phylogeographic data sets". In: *Molecular Ecology* 24.6 (2015), pp. 1164–1171. DOI: 10.1111/mec.13108.

[92] Simon Y W Ho, Beth Shapiro, Matthew J. Phillips, Alan Cooper, and Alexei J. Drummond. "Evidence for Time Dependency of Molecular Rate Estimates". In: *Systematic Biology* 56.June (2007), pp. 515–522.

[93] David Penny. "Relativity for molecular clocks". In: *Nature* 436.July (2005), pp. 183–184. DOI: 10.1038/436183a.

[94] Simon Y W Ho, Matthew J. Phillips, Alan Cooper, and Alexei J. Drummond. "Time dependency of molecular rate estimates and systematic overestimation of recent divergence times". In: *Molecular Biology and Evolution* 22.7 (2005), pp. 1561–1568. DOI: 10.1093/molbev/msi145.

[95] Richard D. Gregory and Arco van Strien. "Wild bird indicators : using composite population trends of". In: *Ornithological Science* 22.1 (2010), pp. 3–22. DOI: 10.2326/osj.9.3.

[96] Henri Weimerskirch, Pablo Inchausti, Christophe Guinet, and Christophe Barbraud. "Trends in bird and seal populations as indicators of a system shift in the Southern Ocean". In: *Antarctic Science* 15.2 (2003), pp. 249–256. DOI: 10.1017/S0954102003001202.

[97] Pertti Koskimies. "Birds as a tool in environmental monitoring". In: *Annales Zoologici Fennici* 26.3 (1989), pp. 153–166.

[98] Martin J. Westgate, Ayesha I. T. Tulloch, Philip S. Barton, Jennifer C. Pierson, and David B. Lindenmayer. "Optimal taxonomic groups for biodiversity assessment: A meta-analytic approach". In: *Ecography* February (2016), pp. 1–10. DOI: 10.1111/ecog.02318.

[99] P. Zion Klos, John T. Abatzoglou, Alycia Bean, et al. "Indicators of Climate Change in Idaho: An Assessment Framework for Coupling Biophysical Change and Social Perception a". In: *Weather, Climate, and Society* 7.3 (July 2015), pp. 238–254. DOI: 10.1175/WCAS-D-13-00070.1.

[100] Timothy J O'Connell, Laura E Jackson, and Robert P Brooks. "Bird Guilds as Indicators of Ecological Condition in the Central Appalachians". In: *Ecological Applications* 10.106 (2000), pp. 1706–1721. DOI: 10.1890/1051-0761(2000)010.

[101] Richard Inger, Richard Gregory, James P. Duffy, Iain Stott, Petr Vorisek, and Kevin J. Gaston. "Common European birds are declining rapidly while less abundant species' numbers are rising". In: *Ecology Letters* 18.1 (2015), pp. 28–36. DOI: 10.1111/ele.12387.

[102] Sergi Herrando, Marc Anton, Francesc Sardà-Palomera, Gerard Bota, Richard D. Gregory, and Lluís Brotons. "Indicators of the impact of land use changes using large-scale bird surveys: Land abandonment in a Mediterranean region". In: *Ecological Indicators* 45 (2014), pp. 235–244. DOI: 10.1016/j.ecolind.2014.04.011.

[103] Philip A Stephens, Lucy R Mason, Rhys E Green, Richard D Gregory, John R Sauer, Jamie Alison, Ainars Aunins, and Lluís Brotons. "Consistent response of bird populations to climate change on two continents". In: *Science (New York, N.Y.)* 352.6281 (2016), pp. 84–87.

[104] Frank T. Burbrink, Yvonne L. Chan, Edward A. Myers, Sara Ruane, Brian Tilston Smith, and Michael J. Hickerson. "Asynchronous demographic responses to Pleistocene climate change in Eastern Nearctic vertebrates". In: *Ecology Letters* (2016). DOI: 10.1111/ele.12695.

[105] Jeff J. Shi and Daniel L. Rabosky. "Speciation dynamics during the global radiation of extant bats". In: *Evolution* 69.6 (June 2015), pp. 1528–1545. DOI: 10.1111/evo.12681.

[106] Jean Pierre Bocquet-Appel and Ofer Bar-Yosef. "The neolithic demographic transition and its consequences". In: *The Neolithic Demographic Transition and its Consequences* March (2008), pp. 1–542. DOI: 10.1007/978-1-4020-8539-0.

[107] Hong-xiang Zheng, Shi Yan, Zhen-dong Qin, Yi Wang, Jing-ze Tan, Hui Li, and Li Jin. "Major Population Expansion of East Asians Began before Neolithic Time: Evidence of mtDNA Genomes". In: *PLoS ONE* 6.10 (Oct. 2011). Ed. by Toomas Kivisild, e25835. DOI: 10.1371/journal.pone.0025835.

[108] Hong-xiang Zheng, Shi Yan, Zhen-dong Qin, and Li Jin. "MtDNA analysis of global populations support that major population expansions began before Neolithic Time". In: *Scientific Reports* 2.1 (Dec. 2012), p. 745. DOI: 10.1038/srep00745.

[109] Carla Aimé, Guillaume Laval, Etienne Patin, et al. "Human Genetic Data Reveal Contrasting Demographic Patterns between Sedentary and Nomadic Populations That Predate the Emergence of Farming". In: *Molecular Biology and Evolution* 30.12 (Dec. 2013), pp. 2629–2644. DOI: 10.1093/molbev/mst156.

[110] Pierpaolo Maisano Delser, Rita Neumann, Stéphane Ballereau, Pille Hallast, Chiara Batini, Daniel Zadik, and Mark A. Jobling. "Signatures of human European Palaeolithic expansion shown by resequencing of non-recombining X-chromosome segments". In: *European Journal of Human Genetics* 25.4 (2017), pp. 485–492. DOI: 10.1038/ejhg.2016.207.

[111] Phillip Endicott, Simon Y W Ho, Mait Metspalu, and Chris Stringer. "Evaluating the mitochondrial timescale of human evolution". In: *Trends in Ecology and Evolution* 24.9 (2009), pp. 515–521. DOI: 10.1016/j.tree.2009.04.006.

[112] Adrien Rieux, Anders Eriksson, Mingkun Li, et al. "Improved Calibration of the Human Mitochondrial Clock Using Ancient Genomes". In: *Molecular Biology and Evolution* 31.10 (Oct. 2014), pp. 2780–2792. DOI: 10.1093/molbev/msu222.

[113] The 1000 Genomes Project Consortium. "A global reference for human genetic variation". In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. DOI: 10.1038/nature15393.

[114] Alexei J Drummond and Andrew Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC Evolutionary Biology* 7.1 (2007), p. 214. DOI: 10.1186/1471-2148-7-214.

[115] Remco Bouckaert, Joseph Heled, Denise Kühnert, et al. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis". In: *PLoS Computational Biology* 10.4 (2014), pp. 1–6. DOI: 10.1371/journal.pcbi.1003537.

[116] Quentin D Atkinson, Russell D Gray, and Alexei J Drummond. "mtDNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory". In: *Molecular Biology and Evolution* 25.2 (Jan. 2008), pp. 468–474. DOI: 10.1093/molbev/msm277.

[117] Sijia Wang, Nicolas Ray, Winston Rojas, et al. "Geographic patterns of genome admixture in latin American mestizos". In: *PLoS Genetics* 4.3 (2008), pp. 1–9. DOI: 10.1371/journal.pgen.1000037.

[118] Joshua Mark Galanter, Juan Carlos Fernandez-Lopez, Christopher R. Gignoux, et al. "Development of a panel of genome-wide ancestry informative markers to study admixture throughout the americas". In: *PLoS Genetics* 8.3 (2012). DOI: 10.1371/journal.pgen.1002554.

[119] Sonja Meyer, Gunter Weiss, and Arndt von Haeseler. "Pattern of nucleotide susbtitution and rate hterogeneity in the hypervaraible regions I and II of human mtDNA". In: *Genetics* 152 (1999), pp. 1103–1110.

[120] Robert Lanfear, Brett Calcott, Simon Y W Ho, and Stephane Guindon. "PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses". In: *Molecular Biology and Evolution* 29.6 (June 2012), pp. 1695–1701. DOI: 10.1093/molbev/mss020.

[121] Richard R. Hudson, Montgomery Slatkin, and Wayne P. Maddison. "Estimation of levels of gene flow from DNA sequence data". In: *Genetics* 132.2 (1992), pp. 583–589. DOI: PMC1205159.

[122] Antti Sajantila, Abdel-Halim Salem, Peter Savolainen, Karin Bauer, Christian Gierig, and Svante Paabo. "Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population." In: *Proceedings of the National Academy of Sciences* 93.21 (Oct. 1996), pp. 12035–12039. DOI: 10.1073/pnas.93.21.12035.

[123] Oscar Lao, Timothy T Lu, Michael Nothnagel, et al. "Correlation between Genetic and Geographic Structure in Europe". In: *Current Biology* 18.16 (Aug. 2008), pp. 1241–1248. DOI: 10.1016/j.cub.2008.07.049.

[124] Mari Nelis, Tõni Esko, Reedik Mägi, et al. "Genetic structure of europeans: A view from the north-east". In: *PLoS ONE* 4.5 (2009). DOI: 10.1371/journal.pone.0005472.

[125] Anu M. Neuvonen, Mikko Putkonen, Sanni Översti, Tarja Sundell, Päivi Onkamo, Antti Sajantila, and Jukka U. Palo. "Vestiges of an ancient border in the contemporary genetic diversity of North-Eastern Europe". In: *PLoS ONE* 10.7 (2015), pp. 1–19. DOI: 10.1371/journal.pone.0130331.

[126] Roger Ballard. "The South Asian Presence in Britain and its Transnational Connections". In: *Culture and Economy in the Indian Diaspora* (2002), pp. 1–30. DOI: 10.4324/9780203398296.

[127] Marina Silva, Marisa Oliveira, Daniel Vieira, et al. "A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals". In: *BMC Evolutionary Biology* 17.1 (2017), p. 88. DOI: 10.1186/s12862-017-0936-9.

[128] Carla Aimé and Frédéric Austerlitz. "Different kinds of genetic markers permit inference of Paleolithic and Neolithic expansions in humans". In: *European Journal of Human Genetics* December 2016 (2016), pp. 360–365. DOI: 10.1038/ejhg.2016.191.

[129] Monika Karmin, Lauri Saag, Mário Vicente, et al. "A recent bottleneck of Y chromosome diversity coincides with a global change in culture". In: *Genome Research* 25.4 (Apr. 2015), pp. 459–466. DOI: 10.1101/gr.186684.114.

[130] Sarah A Tishkoff, Floyd A Reed, Françoise R Friedlaender, et al. "The Genetic Structure and History of Africans and African Americans". In: *Science* 324.5930 (2009), pp. 1035–1044. DOI: 10.1126/science.1172257.The.

[131] David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L Price, and Lalji Singh. "Reconstructing Indian population history". In: *Nature* 461.7263 (2009), pp. 489–494. DOI: 10.1038/nature08365.

[132] Marta D. Costa, Joana B. Pereira, Maria Pala, et al. "A substantial prehistoric european ancestry amongst ashkenazi maternal lineages". In: *Nature Communications* 4 (2013), pp. 1–10. DOI: 10.1038/ncomms3543.

[133] Martin Richards, Vincent Macaulay, Eileen Hickey, et al. "Tracing European Founder Lineages in the Near Eastern mtDNA Pool". In: *The American Journal of Human Genetics* 67.5 (Nov. 2000), pp. 1251–1276. DOI: 10.1086/321197.

[134] Wolfgang Haak, Oleg Balanovsky, Juan J. Sanchez, et al. "Ancient DNA from European early Neolithic farmers reveals their near eastern affinities". In: *PLoS Biology* 8.11 (2010). DOI: 10.1371/journal.pbio.1000536.

[135] Neus Isern and Joaquim Fort. "Modelling the effect of Mesolithic populations on the slowdown of the Neolithic transition". In: *Journal of Archaeological Science* 39.12 (2012), pp. 3671–3676. DOI: 10.1016/j.jas.2012.06.027.

[136] Kumarasamy Thangaraj, Amrita Nandan, Vishwas Sharma, et al. "Deep rooting In-Situ expansion of mtDNA haplogroup R8 in South Asia". In: *PLoS ONE* 4.8 (2009), pp. 2–8. DOI: 10.1371/journal.pone.0006545.

[137] John R Pannell. "Coalescence in a Metapopulation with Recurrent Local Extinction and Recolonzation". In: *Evolution* 57.5 (May 2003), pp. 949–961. DOI: 10.1111/j.0014-3820.2003.tb00307.x.

[138] Adam Nadachowski. "Origin and History of the Present Rodent Fauna in Poland Based on Fossil Evidence". In: *Acta Theriologica* 34 (1989).

[139] Robert Sommer and Norbert Benecke. "Late-Pleistocene and early Holocene history of the canid fauna of Europe (Canidae)". In: *Mammalian Biology* 70.4 (July 2005), pp. 227–241. DOI: 10.1016/j.mambio.2004.12.001.

[140] Joanna L. Elson, Richard M. Andrews, Patrick F. Chinnery, Robert N. Lightowlers, Douglass M. Turnbull, and Neil Howell. "Analysis of European mtDNAs for Recombination". In: *The American Journal of Human Genetics* 68.1 (Jan. 2001), pp. 145–153. DOI: 10.1086/316938.

[141] Richard E Giles, Hugues Blanc, Howard M Cann, and Douglas C Wallace. "Maternal inheritance of human mitochondrial DNA." In: *Proceedings of the National Academy of Sciences of the United States of America* 77.11 (1980), pp. 6715–9.

[142] Robert M. Zink and George F. Barrowclough. "Mitochondrial DNA under siege in avian phylogeography". In: *Molecular Ecology* 17.9 (2008), pp. 2107–2121. DOI: 10.1111/j.1365-294X.2008.03737.x.

[143] Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. "GenBank". In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D67–D72. DOI: 10.1093/nar/gkv1276.

[144] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, et al. "Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44.D1 (2016), pp. D733–D745. DOI: 10.1093/nar/gkv1189.

[145] Cédric Notredame, Desmond G. Higgins, and Jaap Heringa. "T-coffee: A novel method for fast and accurate multiple sequence alignment". In: *Journal of Molecular Biology* 302.1 (2000), pp. 205–217. DOI: 10.1006/jmbi.2000.4042.

[146] Robert C. Edgar. "MUSCLE: A multiple sequence alignment method with reduced time and space complexity". In: *BMC Bioinformatics* 5 (2004), pp. 1–19. DOI: 10.1186/1471-2105-5-113.

[147] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." In: *Nucleic acids research* 30.14 (2002), pp. 3059–3066. DOI: 10.1093/nar/gkf436.

[148] Mark A. Larkin, Gordon Blackshields, Nigel P. Brown, et al. "Clustal W and Clustal X version 2.0". In: *Bioinformatics* 23.21 (Nov. 2007), pp. 2947–2948. DOI: 10.1093/bioinformatics/btm404.

[149] Ulrich Bodenhofer, Enrico Bonatesta, Christoph Horejš-Kainrath, and Sepp Hochreiter. "Msa: An R package for multiple sequence alignment". In: *Bioinformatics* 31.24 (2015), pp. 3997–3999. DOI: 10.1093/bioinformatics/btv494.

[150] Alexei J Drummond and Remco R Bouckaert. *Bayesian evolutionary analysis with BEAST*. Cambridge University Press, 2015.

[151] Lounès Chikhi, Vitor C. Sousa, Pierre Luisi, Benoit Goossens, and Mark A. Beaumont. "The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes". In: *Genetics* 186.3 (2010), pp. 983–995. DOI: 10.1534/genetics.110.118661.

[152] Thomas Städler, Bernhard Haubold, Carlos Merino, Wolfgang Stephan, and Peter Pfaffelhuber. "The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations". In: *Genetics* 182.1 (2009), pp. 205–216. DOI: 10.1534/genetics.108.094904.

[153] Simon Y. W. Ho, Robert Lanfear, Matthew J. Phillips, Ian Barnes, Jessica A. Thomas, Sergios-Orestis Kolokotronis, and Beth Shapiro. "Bayesian Estimation of Substitution Rates from Ancient DNA Sequences with Low Information Content". In: *Systematic Biology* 60.3 (2011), pp. 366–375. DOI: 10.1093/sysbio/syq099.

[154] Emmanuel Paradis. "Pegas: An R package for population genetics with an integrated-modular approach". In: *Bioinformatics* 26.3 (2010), pp. 419–420. DOI: 10.1093/bioinformatics/btp696.

[155] Michelle M. McMahon and Michael J. Sanderson. "Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes". In: *Systematic Biology* 55.5 (2006), pp. 818–836. DOI: 10.1080/10635150600999150.

[156] Benoit Nabholz, Robert Lanfear, and Jerome Fuchs. "Body mass-corrected molecular rate for bird mitochondrial DNA". In: *Molecular Ecology* 25.18 (2016), pp. 4438–4449. DOI: 10.1111/mec.13780.

[157] Richèl J. C. Bilderbeek and Rampal S. Etienne. "babette : BEAUti 2, BEAST2 and Tracer for R". In: *Methods in Ecology and Evolution* 9.9 (Sept. 2018). Ed. by Michael Matschiner, pp. 2034–2040. DOI: 10.1111/2041-210X.13032.

[158] Daniel L. Ayres, Aaron Darling, Derrick J. Zwickl, et al. "BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics". In: *Systematic Biology* 61.1 (Jan. 2012), pp. 170–173. DOI: 10.1093/sysbio/syr100.

[159] Alexei J. Drummond, Marc A. Suchard, Dong Xie, and Andrew Rambaut. "Bayesian phylogenetics with BEAUti and the BEAST 1.7". In: *Molecular Biology and Evolution* 29.8 (2012), pp. 1969–1973. DOI: 10.1093/molbev/mss075.

[160] Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. "Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7." In: *Systematic biology* 67.5 (2018), pp. 901–904. DOI: 10.1093/sysbio/syy032.

[161] Andrew G. Hope, Simon Y W Ho, Jason L. Malaney, Joseph A. Cook, and Sandra L. Talbot. "Accounting for rate variation among lineages in comparative demographic analyses". In: *Evolution* 68.9 (2014), pp. 2689–2700. DOI: 10.1111/evo.12469.

[162] Alexandre Antonelli, Hannes Hettling, Fabien L. Condamine, et al. "Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of Taxa". In: *Systematic Biology* 66.2 (2017), pp. 153–166. DOI: 10.1093/sysbio/syw066.

[163] Stephen A. Smith, Jeremy M. Beaulieu, and Michael J. Donoghue. "Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches". In: *BMC Evolutionary Biology* 9.1 (2009), p. 37. DOI: 10.1186/1471-2148-9-37.

[164] Richard J. Twitchett. "The palaeoclimatology, palaeoecology and palaeoenvironmental analysis of mass extinction events". In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 232.2-4 (2006), pp. 190–213. DOI: 10.1016/j.palaeo.2005.05.019.

[165] Jens-Christian Svenning, Wolf L. Eiserhardt, Signe Normand, Alejandro Ordonez, and Brody Sandel. "The Influence of Paleoclimate on Present-Day Patterns in Biodiversity and Ecosystems". In: *Annual Review of Ecology, Evolution, and Systematics* 46.1 (Dec. 2015), pp. 551–572. DOI: 10.1146/annurev-ecolsys-112414-054314.

[166] Bin Lu, Yuchi Zheng, Robert W. Murphy, and Xiaomao Zeng. "Coalescence patterns of endemic Tibetan species of stream salamanders (Hynobiidae: Batrachuperus)". In: *Molecular Ecology* 21.13 (2012), pp. 3308–3324. DOI: 10.1111/j.1365-294X.2012.05606.x.

[167] Thomas M. Vignaud, Jeffrey A. Maynard, Raphael Leblois, et al. "Genetic structure of populations of whale sharks among ocean basins and evidence for their historic rise and recent decline". In: *Molecular Ecology* (2014). DOI: 10.1111/mec.12754.

[168] Luciano Calderón, Leonardo Campagna, Thomas Wilke, et al. "Genomic evidence of demographic fluctuations and lack of genetic structure across flyways in a long distance migrant, the European turtle dove". In: *BMC Evolutionary Biology* 16.1 (Dec. 2016), p. 237. DOI: 10.1186/s12862-016-0817-7.

[169] Andrew D. Foote, Kristin Kaschner, Sebastian E. Schultze, et al. "Ancient DNA reveals that bowhead whale lineages survived Late Pleistocene climate change and habitat shifts". In: *Nature Communications* 4 (2013), p. 1677. DOI: 10.1038/ncomms2714.

[170] Eline D Lorenzen, David Nogués-Bravo, Ludovic Orlando, et al. "Species-specific responses of LateQuaternary megafauna to climate and humans". In: *Nature* 479.7373 (2012), pp. 359–364. DOI: 10.1038/nature10574.

[171] Jane Elith and John R. Leathwick. "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time". In: *Annual Review of Ecology, Evolution, and Systematics* 40.1 (2009), pp. 677–697. DOI: 10.1146/annurev.ecolsys.110308.120159.

[172] Damien A. Fordham, H. Resit Akçakaya, Miguel B. Araújo, et al. "Plant extinction risk under climate change: are forecast range shifts alone a good indicator of species vulnerability to global warming?" In: *Global Change Biology* 18.4 (Apr. 2012), pp. 1357–1371. DOI: 10.1111/j.1365-2486.2011.02614.x.

[173] Edward F. Connor, Aaron C. Courtney, and James M. Yoder. "Individuals-Area Relationships: The Relationship between Animal Population Density and Area". In: *Ecology* 81.3 (Mar. 2000), p. 734. DOI: 10.2307/177373.

[174] Eleanor F. Miller, Andrea Manica, and William Amos. "Global demographic history of human populations inferred from whole mitochondrial genomes". In: *Royal Society Open Science* 5.8 (2018), p. 180543. DOI: 10.1098/rsos.180543.

[175] Judy R.M. Allen, Thomas Hickler, Joy S. Singarayer, Martin T. Sykes, Paul J. Valdes, and Brian Huntley. "Last glacial vegetation of Northern Eurasia". In: *Quaternary Science Reviews* 29.19-20 (2010), pp. 2604–2618. DOI: 10.1016/j.quascirev.2010.05.031.

[176] Cedric J. Van Meerbeeck, Hans Renssen, and Didier M. Roche. "How did Marine Isotope Stage 3 and Last Glacial Maximum climates differ? - Perspectives from equilibrium simulations". In: *Climate of the Past* 5.1 (2009), pp. 33–51. DOI: 10.5194/cp-5-33-2009.

[177] Wilfried Thuiller, Damien Georges, Robin Engler, and Frank Breiner. *biomod2: Ensemble Platform for Species Distribution Modeling*. 2019.

[178] BirdLife International and Handbook of the Birds of the World. *Bird species distribution maps of the world. Version 2018.1*. 2018.

[179] Robert J. Hijmans. "Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model". In: *Ecology* 93.3 (2012), pp. 679–688. DOI: 10.1890/11-0826.1.

[180] Robert Bagchi, Mike Crosby, Brian Huntley, et al. "Evaluating the effectiveness of conservation site networks under climate change: Accounting for uncertainty". In: *Global Change Biology* 19.4 (2013), pp. 1236–1248. DOI: 10.1111/gcb.12123.

[181] David R. Roberts, Volker Bahn, Simone Ciuti, et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure". In: *Ecography* 40.8 (2017), pp. 913–929. DOI: 10.1111/ecog.02881.

[182] Joanne M. Potts and Jane Elith. "Comparing species abundance models". In: *Ecological Modelling* 199.2 (2006), pp. 153–163. DOI: 10.1016/j.ecolmodel.2006.05.025.

[183] Christine Howard, Philip A. Stephens, James W. Pearce-Higgins, Richard D. Gregory, and Stephen G. Willis. "Improving species distribution models: The value of data on abundance". In: *Methods in Ecology and Evolution* 5.6 (2014), pp. 506–513. DOI: 10.1111/2041-210X.12184.

[184] Alison Johnston, Daniel Fink, Mark D. Reynolds, et al. "Abundance models improve spatial and temporal prioritization of conservation resources". In: *Ecological Applications* 25.7 (Oct. 2015), pp. 1749–1756. DOI: 10.1890/14-1826.1. arXiv: 9809069v1 [gr-qc].

[185] Tom Auer, Daniel Fink, and Matthew Strimas-Mackey. *ebirdst: Tools for loading, plotting, mapping and analysis of eBird Status and Trends data products. R package version 0.1.0*. 2019.

[186] Neil Howell, Christy Bogolin Smejkal, D.A. Mackey, P.F. Chinnery, D.M. Turnbull, and Corinna Herrnstadt. "The Pedigree Rate of Sequence Divergence in the Human Mitochondrial Genome: There Is a Difference Between Phylogenetic and Pedigree Rates". In: *The American Journal of Human Genetics* 72.3 (2003), pp. 659–670. DOI: 10.1086/368264.

[187] Simon Y M Ho. "Calibrating molecular estimates of substitution rates and divergence times in birds". In: *Journal of Avian Biology* 38.4 (2007), pp. 409–414. DOI: 10.1111/j.2007.0908-8857.04168.x.

[188] Jed O. Kaplan. "Wetlands at the Last Glacial Maximum: Distribution and methane emissions". In: *Geophysical Research Letters* 29.6 (2002), pp. 3–1. DOI: 10.1029/2001GL013366.

[189] Paul J. Valdes, David J. Beerling, and Colin E. Jonhson. "The ice age methane budget". In: *Geophysical Research Letters* 32.2 (2005), pp. 1–4. DOI: 10.1029/2004GL021004.

[190] Peter M. Lafleur. "Connecting Atmosphere and Wetland: Energy and Water Vapour Exchange". In: *Geography Compass* 2.4 (2008), pp. 1027–1057. DOI: 10.1111/j.1749-8198.2007.00132.x.

[191] Ying Fan and Gonzalo Miguez-Macho. "A simple hydrologic framework for simulating wetlands in climate and earth system models". In: *Climate Dynamics* 37.1 (2011), pp. 253–278. DOI: 10.1007/s00382-010-0829-8.

[192] Judy R. M. Allen, Ute Brandt, Achim Brauer, et al. "Rapid environmental changes in southern Europe during the last glacial period". In: *Nature* 400.6746 (Aug. 1999), pp. 740–743. DOI: 10.1038/23432.

[193] Karel H Voous. "List of recent Holarctic bird species". In: *Ibis* (1977).

[194] Sudhir Kumar, Glen Stecher, and Koichiro Tamura. "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets". In: *Molecular biology and evolution* 33.7 (2016), pp. 1870–1874. DOI: 10.1093/molbev/msw054.

[195] Julie D. Thompson, Desmond G Higgins, and Toby J Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". In: *Nucleic Acids Research* 22.22 (1994), pp. 4673–4680. DOI: 10.1093/nar/22.22.4673.

[196] Jessica W. Leigh and David Bryant. "Popart: Full-Feature Software for Haplotype Network Construction". In: *Methods in Ecology and Evolution* 6.9 (2015), pp. 1110–1116. DOI: 10.1111/2041-210X.12410.

[197] John Dunning. *CRC Handbook of Avian Body Masses, Second Edition.* 2nd ed. CRC Press., 2007, p. 672.

[198] Remco Bouckaert and Alexei Drummond. "b{M}odel{T}est: {B}ayesian phylogenetic site model averaging and model comparison". In: *bioRxiv* ii (2015), p. 20792. DOI: 10.1101/020792.

[199] Walter Jetz, Gavin H. Thomas, Jeffery B. Joy, Klaas Hartmann, and Arne O. Mooers. "The global diversity of birds in space and time". In: *Nature* 491.7424 (2012), pp. 444–448. DOI: 10.1038/nature11631.

[200] Robert Beyer, Mario Krapp, and Andrea Manica. "A systematic comparison of bias correction methods for paleoclimate simulations". In: *Climate of the Past Discussions* February (2019), pp. 1–23. DOI: 10.5194/cp-2019-11.

[201] Matthew E. Aiello-Lammens, Robert A. Boria, Aleksandar Radosavljevic, Bruno Vilela, and Robert P. Anderson. "spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models". In: *Ecography* 38.5 (2015), pp. 541–545. DOI: 10.1111/ecog.01132.

[202] David R.B Stockwell and A.Townsend Peterson. "Effects of sample size on accuracy of species distribution models". In: *Ecological Modelling* 148.1 (Feb. 2002), pp. 1–13. DOI: 10.1016/S0304-3800(01)00388-X.

[203] Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. "blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models". In: *Methods in Ecology and Evolution* 10.2 (2019), pp. 225–232. DOI: 10.1111/2041-210X.13107.

[204] Mauro Galetti, Marcos Moleón, Pedro Jordano, et al. "Ecological and evolutionary legacy of megafauna extinctions". In: *Biological Reviews* 93.2 (2018), pp. 845–862. DOI: 10.1111/brv.12374.

[205] Korbinian Strimmer and Oliver G. Pybus. "Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot". In: *Molecular Biology and Evolution* 18.12 (Dec. 2001), pp. 2298–2305. DOI: 10.1093/oxfordjournals.molbev.a003776.

[206] José Luis Blanco-Pastor, Mario Fernández-Mazuecos, Alberto J. Coello, Julia Pastor, and Pablo Vargas. "Topography explains the distribution of genetic diversity in one of the most fragile European hotspots". In: *Diversity and Distributions* 25.1 (2019), pp. 74–89. DOI: 10.1111/ddi.12836.

[207] Eric de Silva, Neil M. Ferguson, and Christophe Fraser. "Inferring pandemic growth rates from sequence data". In: *J R Soc Interface* 9.73 (2012), pp. 1797–1808. DOI: 10.1098/rsif.2011.0850.

[208] Mark De Bruyn, Brenda L. Hall, Lucas F. Chauke, Carlo Baroni, Paul L. Koch, and A. Rus Hoelzel. "Rapid response of a marine mammal species to holocene climate and habitat change". In: *PLoS Genetics* 5.7 (2009). DOI: 10.1371/journal.pgen.1000554.

[209] Eleftheria Palkopoulou, Love Dalen, Adrain M. Lister, et al. "Holarctic genetic structure and range dynamics in the woolly mammoth". In: *Proc Biol Sci* 280.1770 (2013), p. 20131910. DOI: 10.1098/rspb.2013.1910.

[210] Antonio González-Martín, Amaya Gorostiza, Lucía Regalado-Liu, et al. "Demographic History of Indigenous Populations in Mesoamerica Based on mtDNA Sequence Data". In: *PLOS ONE* 10.8 (Aug. 2015). Ed. by Alessandro Achilli, e0131791. DOI: 10.1371/journal.pone.0131791.

[211] Daniel F Mazerolle, Kevin W Dufour, Keith A Hobson, and Heidi E Den Haan. "Effects of large-scale climatic fluctuations on survival and production of young in a Neotropical migrant songbird, the yellow warbler Dendroica petechia". In: *Journal of Avian Biology* 36.2 (Feb. 2005), pp. 155–163. DOI: 10.1111/j.0908-8857.2005.03289.x.

[212] Rachael A. Bay, Ryan J. Harrigan, Vinh Le Underwood, H. Lisle Gibbs, Thomas B Smith, and Kristen Ruegg. "Genomic signals of selection predict climate-driven population declines in a migratory bird". In: *Science* 359.6371 (Jan. 2018), pp. 83–86. DOI: 10.1126/science.aan4380.

[213] Vera M. Warmuth and Hans Ellegren. "Genotype-free estimation of allele frequencies reduces bias and improves demographic inference from RADSeq data". In: *Molecular Ecology Resources* 19.3 (May 2019), pp. 586–596. DOI: 10.1111/1755-0998.12990.

[214] Rasmus Nielsen, Thorfinn Korneliussen, Anders Albrechtsen, Yingrui Li, and Jun Wang. "SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data". In: *PLoS ONE* 7.7 (2012). DOI: 10.1371/journal.pone.0037558.

[215] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. "ANGSD: Analysis of Next Generation Sequencing Data". In: *BMC Bioinformatics* 15.1 (Dec. 2014), p. 356. DOI: 10.1186/s12859-014-0356-4.

[216] Franck Prugnolle, Andrea Manica, and François Balloux. "Geography predicts neutral genetic diversity of human populations". In: *Current Biology* 15.5 (2005), pp. 159–160. DOI: 10.1016/j.cub.2005.02.038.

[217] Brad H McRae. "Isolation by resistance". In: *Evolution* 60.8 (Aug. 2006), pp. 1551–61.

[218] Brad H. McRae, Brett G. Dickson, Timothy H. Keitt, and Viral B. Shah. "Using circut theory to model connectivity in ecology, evolution, and conservation". In: *Ecology* 89.10 (Oct. 2008), pp. 2712–24. DOI: 10.1890/07-1861.1.

[219] R Core Team. *R: A Language and Environment for Statistical Computing*. 2019.

[220] Anders Eriksson, Lia Betti, Andrew. D. Friend, et al. "Late Pleistocene climate change and the global expansion of anatomically modern humans". In: *Proceedings of the National Academy of Sciences* 109.40 (Oct. 2012), pp. 16089–16094. DOI: 10.1073/pnas.1209494109.

[221] Staffan Bensch, Darren E. Irwin, Jessica H. Irwin, Laura Kvist, and Susanne Åkesson. "Conflicting patterns of mitochondrial and nuclear DNA diversity in Phylloscopus warblers". In: *Molecular Ecology* 15.1 (2006), pp. 161–171. DOI: 10.1111/j.1365-294X.2005.02766.x.

[222] Katalin Csilléry, Olivier François, and Michael G B Blum. "Abc: An R package for approximate Bayesian computation (ABC)". In: *Methods in Ecology and Evolution* 3.3 (2012), pp. 475–479. DOI: 10.1111/j.2041-210X.2011.00179.x.

[223] Emmanuel Milot, H. Lisle Gibbs, and Keith A. Hobson. "Phylogeography and genetic structure of northern populations of the yellow warbler (Dendroica petechia)". In: *Molecular Ecology* 9.6 (2000), pp. 667–681. DOI: 10.1046/j.1365-294X.2000.00897.x.

[224] Benjamin C. Haller and Philipp W. Messer. "SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model". In: *Molecular Biology and Evolution* 36.3 (2019), pp. 632–637. DOI: 10.1093/molbev/msy228.

[225] Mateusz Baca, Adam Nadachowski, Grzegorz Lipecki, et al. "Impact of climatic changes in the late pleistocene on migrations and extinction of mammals in Europe: Four case studies". In: *Geological Quarterly* 61.2 (2017), pp. 291–304. DOI: 10.7306/gq.1319.

[226] Enrico A. Ruiz, Enriqueta Velarde, and Andres Aguilar. "Demographic history of Heermann's Gull (Larus heermanni) from late Quaternary to present: Effects of past climate change in the Gulf of California". In: *The Auk* 134.2 (2017), pp. 308–316. DOI: 10.1642/AUK-16-57.1.

[227] Rasmus Nielsen, Joshua M Akey, Mattias Jakobsson, Jonathan K Pritchard, Sarah Tishkoff, Eske Willerslev, and Wellcome Genome Campus. "Tracing the peopling of the world through genomics". In: *Nature* 541.7637 (2017), pp. 302–310. DOI: 10.1038/nature21347.Tracing.

[228] Christopher Blair, Kellie L Heckman, Amy L Russell, and Anne D Yoder. "Multilocus coalescent analyses reveal the demographic history and speciation patterns of mouse lemur sister species". In: *BMC Evolutionary Biology* 14.1 (2014), p. 57. DOI: 10.1186/1471-2148-14-57.

[229] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. "GenBank". In: *Nucleic Acids Research* 41.D1 (2013), pp. 36–42. DOI: 10.1093/nar/gks1195.

[230] William Stewart Grant and Wei Cheng. "Incorporating deep and shallow components of genetic structure into the management of Alaskan red king crab". In: *Evolutionary Applications* 5.8 (2012), pp. 820–837. DOI: 10.1111/j.1752-4571.2012.00260.x.

[231] Jose R. Alvarez, Dmitry Skachkov, Steven E. Massey, Alan Kalitsov, and Julian P. Velev. "DNA/RNA transverse current sequencing: Intrinsic structural noise from neighboring bases". In: *Frontiers in Genetics* 6.JUN (2015), pp. 1–11. DOI: 10.3389/fgene.2015.00213.

[232] Jared Diamond and Peter Bellwood. "Farmers and Their Languages: The First Expansions". In: *Science* 300.5619 (Apr. 2003), pp. 597–603. DOI: 10.1126/science.1078208.

[233] Mark Collard, Kevan Edinborough, Stephen Shennan, and Mark G. Thomas. "Radiocarbon evidence indicates that migrants introduced farming to Britain". In: *Journal of Archaeological Science* 37.4 (2010), pp. 866–870. DOI: 10.1016/j.jas.2009.11.016.

[234] Lucy J.E. Cramp, Richard P. Evershed, Lavento Mika, et al. "Neolithic dairy farming at the extreme of agriculture in northern Europe". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1791 (2014), p. 20140819. DOI: 10.1098/rspb.2014.0819.

[235] Gloria González-Fortes, Eppie R. Jones, Emma Lightfoot, et al. "Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin". In: *Current Biology* 27.12 (2017), pp. 1801–1810. DOI: 10.1016/j.cub.2017.05.023.

[236] Fabio Silva, Chris J. Stevens, Alison Weisskopf, Cristina Castillo, Ling Qin, Andrew Bevan, and Dorian Q. Fuller. "Modelling the geographical origin of rice cultivation in Asia using the rice archaeological database". In: *PLoS ONE* 10.9 (2015), pp. 1–21. DOI: 10.1371/journal.pone.0137024.

Appendix A

Supplementary Information for Chapter 2

COMPOSITION	NUMBER OF SITES	BEST MODEL	SUBSTITUTION RATE (μ/SITE/YEAR) (UNITS OF 10^{-8})
HVS1+HVS2	1122	TN93+I+G	31.434
R+TRNA	4018	TN93+I+G	1.007
PC1+PC2	7565	TN93+I+G	0.756
PC3	3777	TN93+G	3.323

Table A.1 Composition of best partitioning scheme for the mtDNA (obtained with PartitionFinder software). Substitution model for each partition and substitution rates were as per by Rieux *et al.* [112].

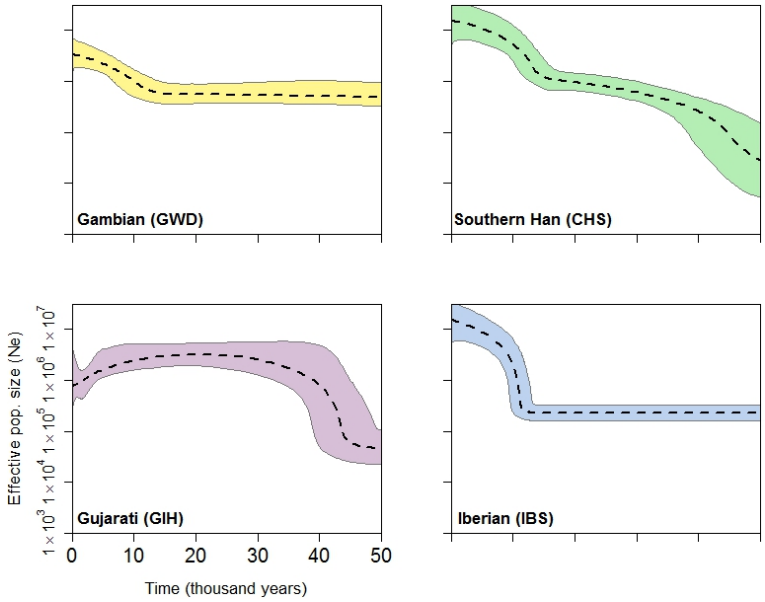


Fig. A.1 EBSPs of one population from each of the four major regions. All plots are on the same scale. Dotted line is the median estimate and the thin grey lines show the boundary of the 95% CPD interval. The populations with the largest sample sizes were selected from each major region. The European IBS population had 107 individuals; East Asia 105 samples from CHS; Africa 113 GWD samples and South Asia 103 samples from GIH. The x-axis represents time from the present in years.

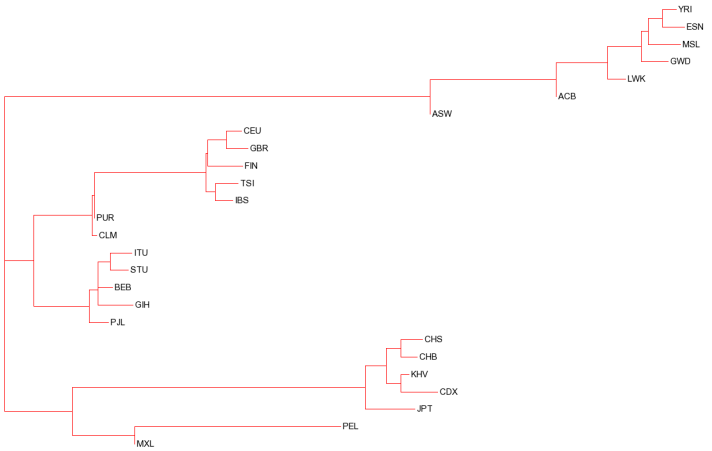


Fig. A.2 Neighbour-joining tree relating populations used from the 1000 Genomes Project, based Fst calculated from autosomal SNPs.

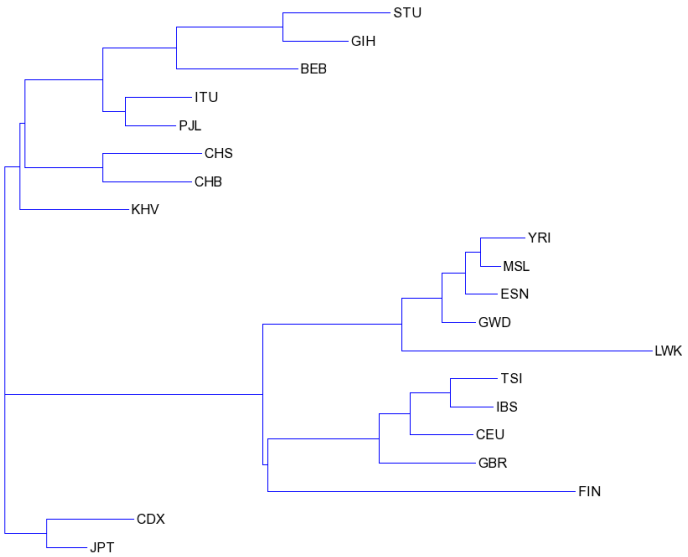


Fig. A.3 Neighbour-joining tree relating populations used from the 1000 Genomes Project, based on the similarity of their Bayesian skyline plot profiles.

Continental group	Pop.	Approx. timing of expansion (years ago)	Median initial Ne	Median current Ne	Agriculture intro.
AFR	YRI	9317	510	9900	Oldest evidence ~5-6kya
	LWK	2076	1000	1450	
	ESN	10953	430	4500	
	MSL	16821	475	2500	
	GWD	15670	450	2400	
EUR	GBR	9598	140	5300	Oldest evidence is ~9kya and more recent in Northern Europe
	FIN	23830	10	600	
	TSI	14978	220	9000	
	CEU	14876	170	5350	
	IBS	12397	230	9000	
SA	ITU	38743	40	2300	Oldest evidence is ~8-9kya
	BEB	40832	15	2800	
	GIH	43214	30	600	
	STU	45141	20	700	
	PJL	40855	35	2800	
EA	JPT	5903	30	5000	Oldest evidence is ~9-10kya
	CHS	14693	16	12500	
	CDX	38349	220	1700	
	CHB	12170	35	19000	
	KHV	14142	17	2400	

Table A.2 Summary table of values for population N_e at 50 kya and present day, timings of greatest population expansion and approximate dates for the introduction of agriculture to each region, taken as the best supported date found in the literature for when agriculture was established. Most of these dates are open to debate. Note, the algorithm used to extract the date of strongest inferred expansion requires data either side and therefore fails to pick up expansions very close to either extreme of a profile. Profiles for East Asian populations often exhibit two expansion phases and only the strongest is recorded. [232, 233, 234, 235, 236, 135]

Appendix B

Supplementary Information for Chapter 3

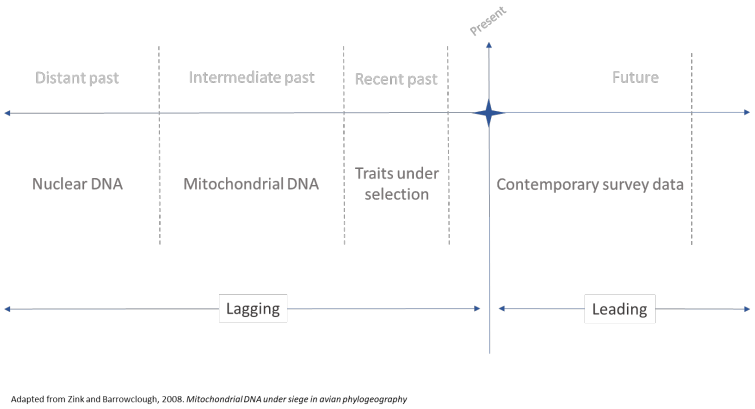
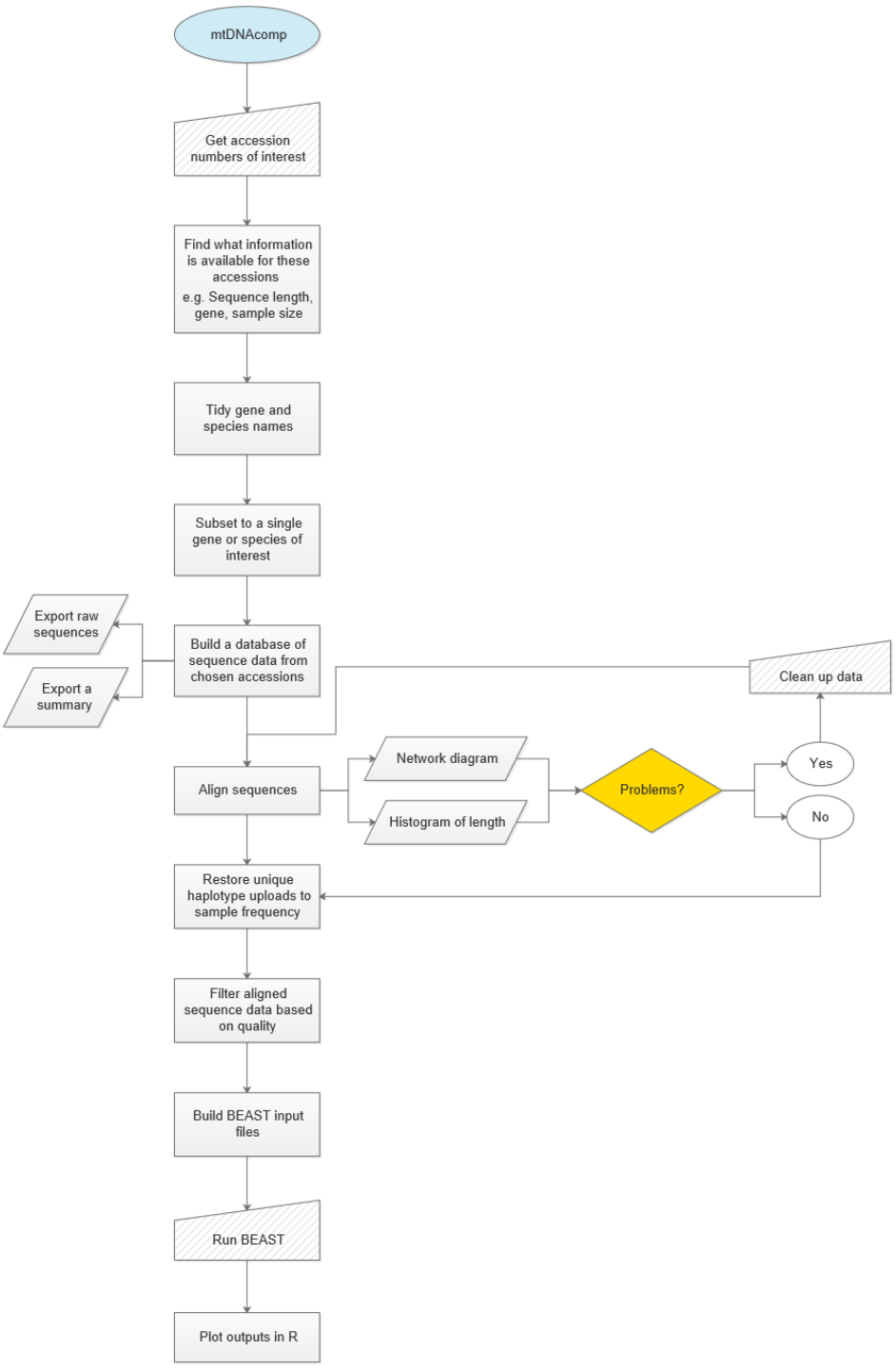


Fig. B.1 Differing sensitivity of a range of loci for reconstructing different periods of population history.

mtDNAcomp Vignette

E.F.Miller
Department of Zoology, University of Cambridge.
em618@cam.ac.uk

23 August 2018


mtDNAcomp
Get accession numbers of interest
Find what information is available for these accessions
e.g. Sequence length, gene, sample size
Tidy gene and species names
Subset to a single gene or species of interest
Export raw sequences
Export a summary
Build a database of sequence data from chosen accessions
Clean up data
Align sequences
Network diagram
Histogram of length
Problems?
Yes
No
Restore unique haplotype uploads to sample frequency
Filter aligned sequence data based on quality
Build BEAST input files
Run BEAST
Plot outputs in R

Using mtDNAcomp

This vignette describes the mtDNAcomp package, an R library designed to support comparative analyses of Bayesian Skyline Plot (BSP) population histories based on mtDNA sequence data from multiple studies.

mtDNAcomp includes functions to retrieve, align, summarise, and maniplulate seuquences downloaded from GenBank, as well as generating basic BEAST2 input files. There are also accessory functions to analyse and plot the outputs of BEAST2 runs.

To install mtdnacomp you will need devtools:

```
install.packages("devtools")
library(devtools)
```

Then to load mtDNAcomp:

```
library(mtDNAcomp)
```

Creating input accession number file

To start a project comparing mitochondrial DNA from multiple individuals, species, and studies, we first need to have a list of unique GenBank accessions to explore.

These accessions can be acquired in multiple ways. A simple method would be to undertake a broad search of GenBank, e.g. open the NCBI webpage with code such as below:

```
browseURL( "https://www.ncbi.nlm.nih.gov/nuccore")
```

Set search terms along the lines of "birds"[porgn:__txid8782] in the Nucleotide database with the 'Genetic compartments - Mitochondria' box checked.

Then click the 'Send to:' drop down, check the 'Complete Record' radio button, then under 'Choose Destination:' the 'File' radio button, and finally, under 'Format:' select 'Accession List' then press 'Create File'

The output produced by GenBank is likely to be a '.seq' file, or, if using a list of accessions gathered in a different way (e.g. compiled by hand), a list of accessions saved in a .csv format is okay as long as there are no headers or row names.

For this vignette we will work with a fixed set of accessions from the file "vignette_accessions.csv" from the pacakge. This file can be accessed from the extdata directory through the 'system.file' command (see below for an example of its usage).

An initial sweep of available information

Firstly, we build a dataframe that contains information on all the genes / sequences associated with each accession numbers to explore what information is available.

```
path_to_file <- system.file("extdata","vignette_accessions.csv",
                            package="mtDNAcomp")
GB_data <- build_genbank_df(accession_file_name = path_to_file)
```

GB_data should be 860 observations.

Tidying up raw information

Within GenBank, the same single sequence is often associated to multiple features (e.g. 'source', 'gene', and 'CDS'). This is visible on the website, where the same sequence is found under multiple 'Feature' tabs. This means that the same sequence will also be grabbed multiple times when scraping data from GenBank. To clean the dataset for later analysis, we must remove duplicated entries.

We must also control for the bredth of possible names used to describe a single gene as individual studies/groups/projects upload data to GenBank using a range of possible synonyms, abreviations, and misspellings. The first step is to standadise nomenclture across the dataframe by converting gene names to a user defined set of 'standard' nomenclatures.

By default, the `standardise_gene_names` function loads a file containing alternate abreviations, common misspellings, and other frequent errors for 18 commonly sequenced mitochondiral genes. The user can upload a custom file by specifying the different file as the second variable in the function: `standardise_gene_names(df_to_update, names_to_replace)`

```
GB_data <- standardise_gene_names(GB_data)
```

Then we remove the duplicates.

```
GB_data <- remove_duplicates(df_to_update = GB_data)
nrow(GB_data)
```

```
## [1] 500
```

We do not expect the number of accessions we're exploring and number of observations to match at this stage. Indeed, in this example we see that, from the orignal 335 accession numbers in 'GB_data.csv', the GB_data data frame now has 500 observations. This is because the script captures ALL genes associated with each given accession. Every submission to GenBank recieves a unique accession number but these individual submissions can contain data for anything from a single gene through to whole genome data.

Check what information is available

Firstly, for what genes are there data?

```
GB_genes <- droplevels(as.data.frame(unique(GB_data$gene_name)))
```

These data are still messy.

We can tidy the data a little by removing some gene names that are unlikely to be useable or comparable with other sequences e.g. removing any names that are just numbers, removing names over a certain length, and/or dropping other common unwanted patterns.

```
GB_genes <- as.data.frame(GB_genes[grep("[[:alpha:]]",
                                        GB_genes$`unique(GB_data$gene_name)`), ])
colnames(GB_genes) <- "gene_name"

GB_genes <- as.data.frame(GB_genes[!nchar(as.character(GB_genes$gene_name)) > 20, ])

patt <- c("RNA|unknown|trn|ATP|similar|central|duplicated|tandem|repeated|other|myoglobin")
colnames(GB_genes) <- "gene_name"
GB_genes <- droplevels(as.data.frame(GB_genes[!grepl(patt, x = GB_genes$gene_name), ]))
```

Secondly; For what species are there data?

```
GB_species <- droplevels(as.data.frame(unique(GB_data$sci_nam)))
```

At the moment:

```
## [1] "GB_genes has 13  unique gene names while GB_species has 5 unique species names"
```

However, looking at these data in more detail shows us that there are still some spurious entries being included. For example, GB_species includes both *Motacilla alba* and *Motacilla alba alboides*, a recognised subspecies but, in this instance, data that we want to group with *Motacilla alba* more broadly.

```
##             Unique Names
##           Motacilla alba
##     Picoides tridactylus
##        Calidris maritima
##      Pinicola enucleator
##  Motacilla alba alboides
```

Clean up the species names

As stated previously, the amount of freedom in the formating of descriptive information associated with GenBank submissions means that individual submissions can vary the chosen species name, inlcuding using different levels of detail for taxonomic rank. For example, some studies may use subspecies names where others choose not to. Subspecies recognition is frequently debated and sometimes we may want to group together samples with names that aren't an exact match. A 3 word name, not a 2 word scientific name, is a simple pattern to recognise subspecies and we exploit that here.

The `check_poss_synyms` function returns a list of scientific names that are longer than 2 words and these names will be outputted as a .csv file; "poss_synyms.csv"

```
poss_synyms <- check_poss_synyms(data_file = GB_data)
```

If, after investigation, any of these species names need updating or altering then they can be edited within the .csv file. As long as the edited file is saved in the same format, then the `standardise_spp_names` function will reload and integrate any updated names.

For this example, the edited file has been called "poss_sysnyms_updated.csv".

```
GB_data <- standardise_spp_names(data_file = GB_data,
             new_names_file = system.file("extdata", "poss_synyms_updated.csv",
                                          package="mtDNAcomp"))
```

After updating the species names samples for *Motacilla alba alboides* are now labelled as *Motacilla alba* and, therefore, group together for downstream analysis.

```
##          Unique Names
##        Motacilla alba
##  Picoides tridactylus
##     Calidris maritima
##   Pinicola enucleator
```

Filter dataframe

Different ways of creating the original list of accession numbers result in different types of noise being introduced to the dataframe. To retain only sequence data from relevant genes, the dataframe needs to be filtered using the `gene_of_interest` function.

Here we want to look at the *ND2* gene

```
GB_by_gene <- gene_of_interest(gene = "ND2", data_file = GB_data)
```

```
## [1] "GB_by_gene has 333 while GB_data.csv has 335 rows"
```

The discrepancy between accessions originally inputted and the size of the post-filtering dataset is because some accessions numbers in GenBank represent the same samples and these have been removed in `GB_by_gene`.

The two samples dropped were Reference Sequence (RefSeq) accessions. The RefSeq collection aims to provide a collated and stable set of standard reference sequences for studies to build on. Drawn from genomes already available in GenBank and other community databases, they may duplicate existing accessions and, for our puroses, need to be removed.

Extract the available raw sequence data

By simply using the `get_GB_sequence_data` function and the curated accession list, we can now download raw sequence data associated with the specific gene of interest.

```
GB_with_SeqDat <- get_GB_sequence_data(accessions_of_interest = GB_by_gene,
                      gene = "ND2", new_names_file = "poss_synyms_updated.csv")
```

For this vignette, the GB_with_SeqDat file should now be 333 observations with 8 variables.

```
nrow(GB_with_SeqDat)
## [1] 333
ncol(GB_with_SeqDat)
## [1] 8
```

Store out key data files

At this stage it might be helpful to store summary details on the data as well as keeping all the raw, unaligned, sequence data for each species / gene combination. The `export_details` function writes summary details to .csv files while the `export_sequences` function writes out individual .fasta files for each dataset.

```
export_details(data_file = GB_with_SeqDat)

export_sequences(data_file = GB_with_SeqDat)
```

Aligning the sequence data

We have now managed to generate a set of sequences from multiple species covering one gene. However, these sequences often come from multiple independent studies and frequently differ in the gene region they analyse.

Previously, the `export_sequences` function wrote out a .fasta file of raw, unaligned, sequence data for each species / gene combination, starting the file name with the regular expression 'FOR_ALIGNMENT'. We exploit this pattern to capture the list of file names to explore.

```
alignment_files <- list.files(pattern="FOR_ALIGNMENT")
```

For each species, the sequence data needs to be aligned so that we can capture comparable regions of the genome common to each sample. This is done within the `align_and_summarise` function using the ClustalW algorithm, removing any columns with blanks or ambiguos calls.

```
align_and_summarise(alignment_files = alignment_files, max_haps_found_together = 2,
                    minbp = 200)
```

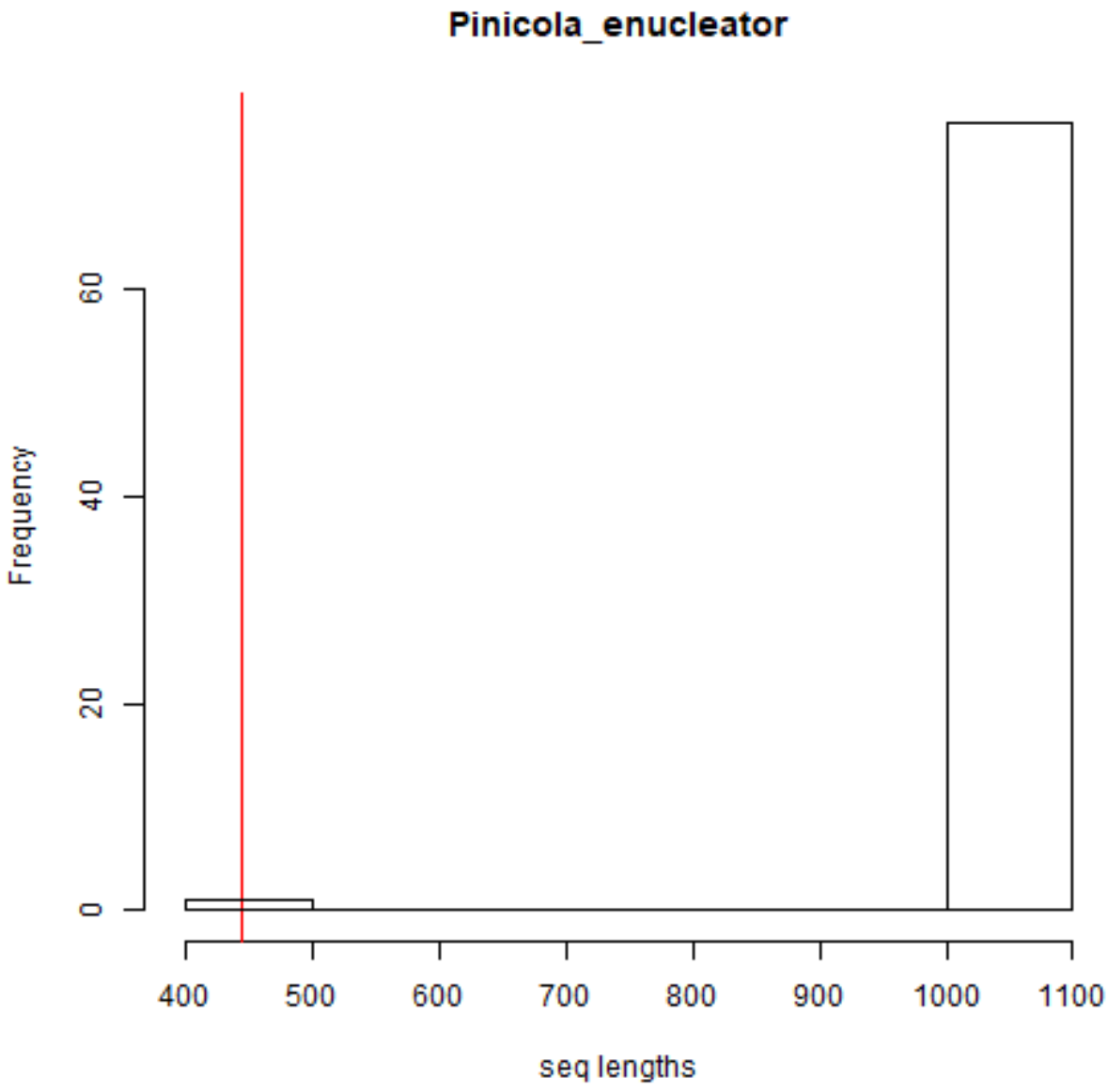
```
## use default substitution matrix
## use default substitution matrix
## use default substitution matrix

## use default substitution matrix
```

Diagnositc plots

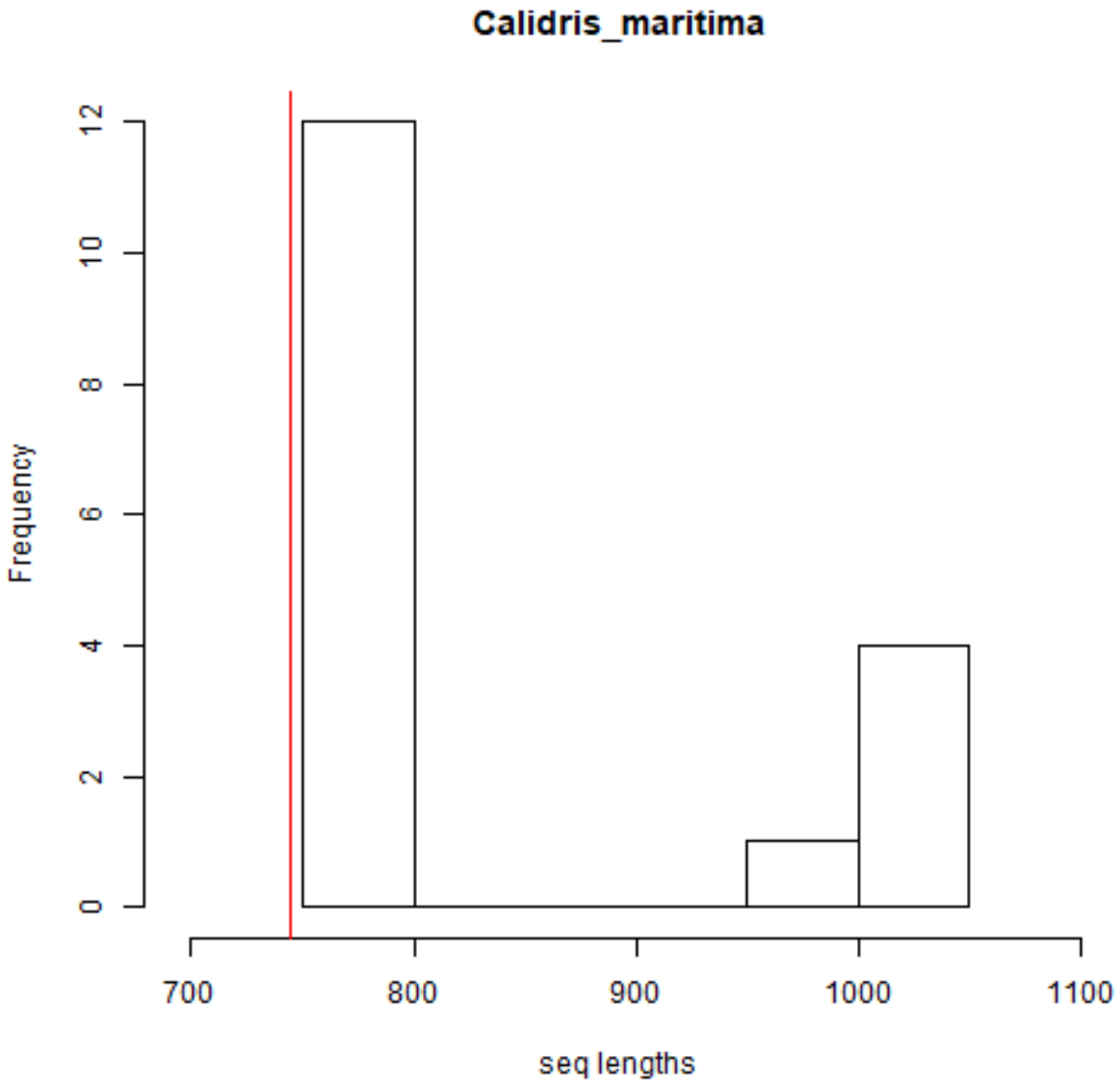
Histogram

Depending on the quality/consistency of the raw sequence data, this step can result in a dramatic reduction in the number of base pairs left in the DNA string. For example, where one or two sequences are very short, or the section of the genome sequenced is different, the overlap between data from separate studies can be very small. In some instances, removal of one or two sequences before alignment could result in a more informative data set.

The impact of the alignment/trimming process is summarised in a diagnostic histogram plot, offering a visual way to identify cases where it would be advantageous to look at the raw data in more detail. The histogram bars show frequency and sequence length of raw, unaligned data and the red line shows the length of the aligned sequences after cropping to the longest section common to all samples.



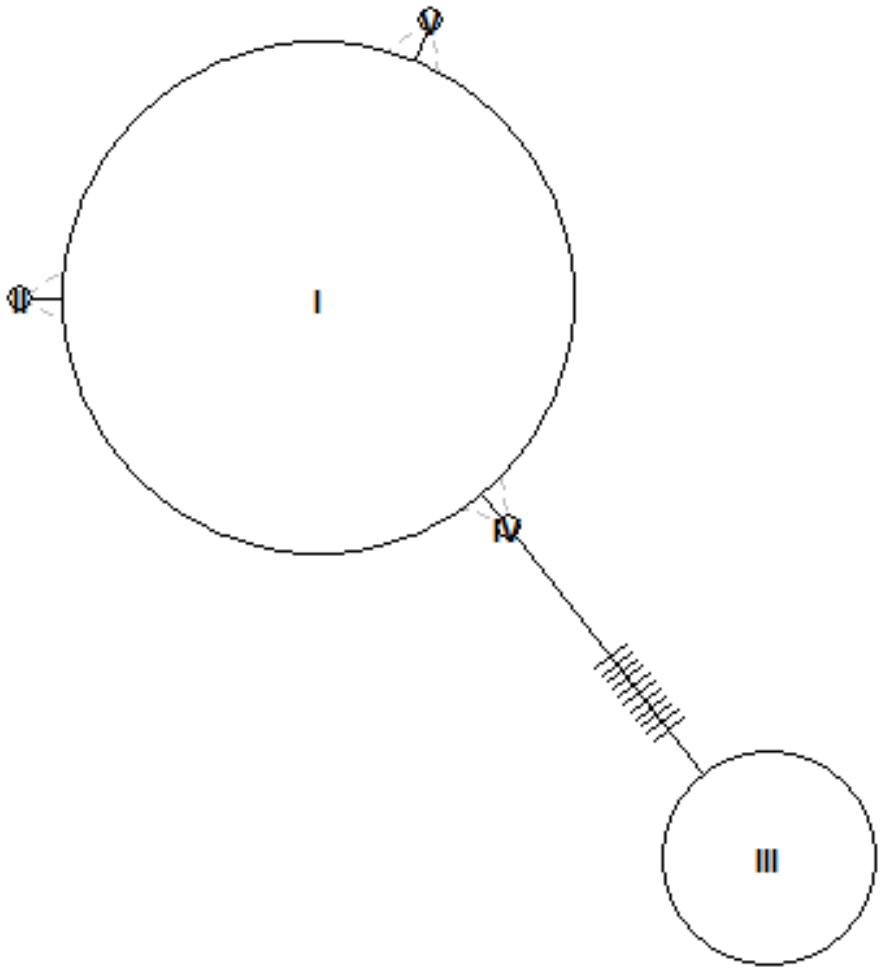
Here we see that, in the *Pinicola enucleator* dataset, the majority of samples have been heavily cropped due to the inclusion of one, shorter, sequence. In this instance, it may be worth reviewing the decision to include the single, much shorter, 450 base pair sample.



Alternatively, the *Calidris maritima* histogram shows that, whilst a few longer sequences have been trimmed by a couple hundered base bairs, the majority of the sequences are being used at nearly full length. This alignment and crop seems good.

Network diagram

The `align_and_summarise` function also produces a haplotype network diagram which helps visuliase the level of structure in a population/sample set.



Here is an example of a network diagram for data from *Picoides tridactylus*. Plots like these help to quickly flag if there are any extreme outliers in the dataset or if the population is heavily structured.

Haplotype frequency

In datasets that include samples from studies which have uploaded a single representative haplotype, instead of creating a new accession for every sample, an aligned sequence output file is not generated. Instead, the papers associated with the unique haplotypes are listed in a .csv file along with the species name; the file is is written out as "More_info_df.csv".

For each of the populations recorded in "More_info_df.csv", the `align_and_summarise` function also creates a file containing each accession number and a frequency column. These files all have the regular pattern "MAGNIFY"" in the file name. Orignal published papers must be tracked down to confirm details of sampling

frequency, new values can then be recorded in the "freq" column. Once updated the file needs to be saved in the same format. In cases where sampling frequency data is not available samples must be excluded.

In this vignette-dataset accessions for *Calidris maritima* and *Motacilla alba* are flagged as needing further investigation. Exploration of the original published papers suggest that data for *Motacilla alba* have indeed been uploaded at sampled frequency and this was essentially a "false alarm". Therefore, we don't want to alter this data so the "MAGNIFY_Motacilla_alba.csv"" file can be left as it is, with a default "freq" column value of "1".

However, exploration of the paper *'A review of the subspecies status of the Icelandic Purple Sandpiper Calidris maritima littoralis'* shows that only unique haplotype sequences were uploaded, rather than a new accession being created for every sample. Therefore, this dataset needs to be manipulated to get to the original sampled frequency.

For the puruposes of this vignette we have created an updated "MAGNIFY_Calidris_maritima.csv" file (found in ../extdata/) which already contains the values for the number of times each haplotype was sampled in the population.

```
magnify_file_list <- list.files(pattern="MAGNIFY")

mag_df <- magnify_to_sampled_freq(magnify_file_list = magnify_file_list)
```

```
## use default substitution matrix

## use default substitution matrix
```

After updating the frequency information the sequences are processed as before - haplotype networks are drawn and .fasta files of the aligned sequence data written out.

To keep accurate summary information of the datasets avalible, we now need to combine the original `info_df` and the newly created `mag_df`. This will give an updated .csv file that contains information on all the data sets we are working with.

```
info_df <- updating_info_df(original_df = "Info_df.csv", new_df = mag_df)
```

Filtering by rules

Data inclusion criteria will vary between studies; there will never be a "one-size-fits-all" set up. Factors such as species life history, species population history, data availablity, data quality, and even broadly the project aims will influence what data are informative.

The following filtering steps are based on a series of rules built around avain mtDNA.

Firstly, we want to drop populations with insufficent sequence data. This includes data with insufficent number of bases, low numbers of haplotypes, low sample size.

```
info_df <- drop_low_sample_size(info_df = info_df, min_sample = 7)

info_df <- drop_low_haplo_number(info_df = info_df, min_haps = 6 )

info_df <- drop_low_sequence_length(info_df = info_df, min_length = 600)
```

After applying these filters we are left with curated datasets from two species. We then want to remove any extreme outliers, considered here to be single samples separated from the nearest haplotype with >30 mutations on a branch. The function `outliers_dropped` writes out an updated version of `info_df` but doesn't return it. Therefore we need to read in the new version from the working directory.

```
what_gets_dropped <- outliers_dropped(max_mutations = 30,info_df = info_df)

info_df <- read.csv("Info_df.csv")

what_gets_dropped
```

```
##      outlier_accession spp_name
## [1,] "EU166960.1"      "new_Motacilla_alba"
```

At this point:

- all the orignal accessions from the accession list have been processed,
- raw sequence data from relevant sections of the genome have been captured,
- sequences have been aligned,
- low resolution/low quality data have been rejected,
- outliers have been removed,
- cleaned sequence data has been written out for use by additional tools.

We want to use these processed data to set up BEAST runs. In order to speed up the process, and reduce the opportunity for human error, we limit the amount of manual set up required by creating basic BEAUti files with the following code. These files will still require some degree of editing in the BEAUti GUI.

The dataset files have been given the prefix "ALIGNED_", making them easy to find. After editing (e.g. dropping outliers), any new versions of aligned data have been given the tag "new_ALIGNED" and should be used in preference to the original files.

```
aligned_files <- list.files(pattern="ALIGNED")


superseeded <- NULL
for(n in 1:length(aligned_files)){
  if (substr(aligned_files[n],1,3)=="new"){
    superseeded <- rbind(gsub("new_",'',aligned_files[n]), superseeded)
  }
}
aligned_files <- aligned_files[!aligned_files%in%superseeded]
```

Build basic xml files

```
setup_basic_xml(gene_name = "ND2", aligned_files = aligned_files)
```

Once created, the .xml files will need to be manually edited. For example, setting up the use of bModelTest - at the time of writing not yet an available option in the `babette` package.

Exploring outputs

Once BEAST runs are completed we need to explore convergence, ESS values, and other metrics.

Here we present a pipeline for handling outputs from the software package Tracer.

An example BEAST .xml input file can be found in ./extdata/ND2_Carpodacus_erythrinus_BEASTinput.xml . After running this file in BEAST v2 4.6 we used Tracerv1 to format output data for export. The resulting file is stored as ./extdata/ND2_Carpodacus_erythrinus_TracerOut.txt

We will use this output file to explore simple plotting/visulisation.

Plotting

A quick look at the structure of the output from Tracer taking only complete rows (NAs can occour at the end of the file but cause issues in later processing).

```
data <-read.table(system.file("extdata","ND2_Carpodacus_erythrinus_TracerOut.txt",
                                    package = "mtDNAcomp"), skip=1, header=T)
data<-data[complete.cases(data),]
head(data)
```

```
##       Time    Mean  Median    Upper   Lower
## 1   0.0000 7952376 7075067 16749959 3933864
## 2 145.5466 7955168 7076337 16713627 3953354
## 3 291.0932 7958742 7079056 16695231 3971612
## 4 436.6397 7958827 7080898 16669858 3975339
## 5 582.1863 7950289 7079365 16601677 3975339
## 6 727.7329 7950633 7082537 16579575 3982642
```

A simple coloured plot can now be created using the code below.

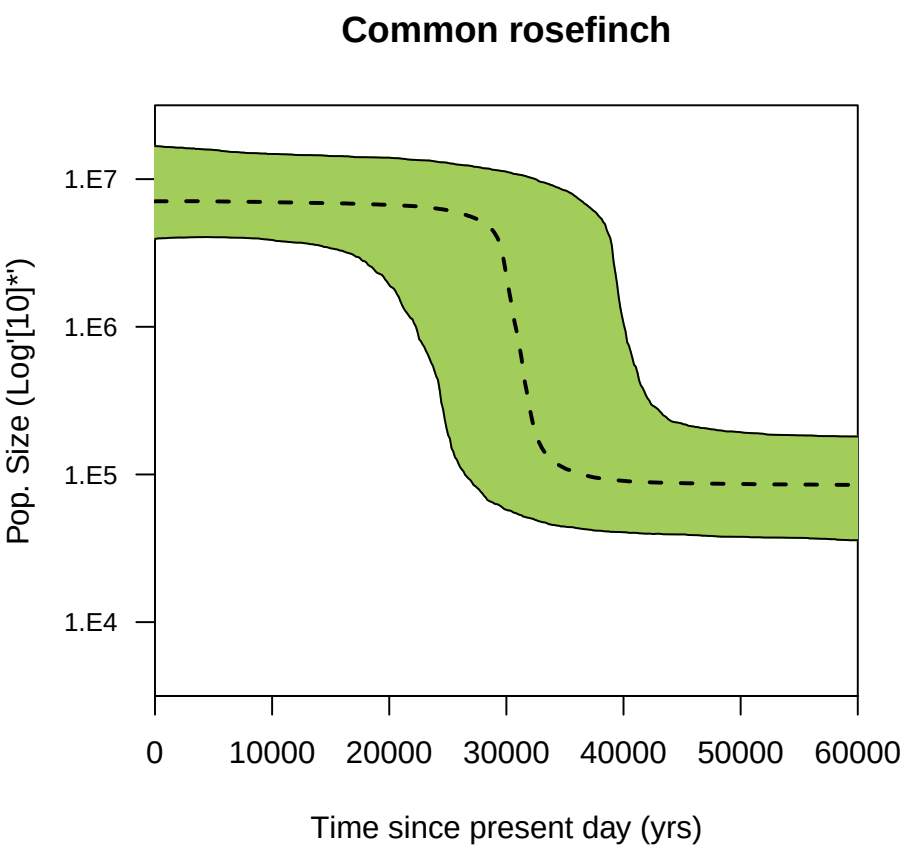
Use the file name as the title

```
file_name <- "ND2_Carpodacus_erythrinus1_TracerOut.txt"
plot_title <- "Common rosefinch"
```

We want to plot the median (log scale) as well as plotting the HPD interval as a coloured polygon.
If analysing data from multiple genes it can be helpful to differentiate the plots on the basis of gene type. Here this is done by colouring the HPD according to the gene.

```
plot(log10(data[,3])~data[,1],type="n",ylim=c(3.5,7.5),xlim=c(0,60000), yaxt="n",
     yaxs="i", ylab = expression("Pop. Size (Log'[10]*')"), xaxs="i",
     xlab = "Time since present day (yrs)")
axis(2, at=c(4,5,6,7), labels = c("1.E4","1.E5","1.E6","1.E7"), las=2, adj=1,
     cex.axis=0.82)


gene <- substr(file_name, start = 1, stop=4)

for(i in 1:length(data[,1])-1) {
  x<-c(data[i,1],data[i+1,1],data[i+1,1],data[i,1])
  y<-log10(c(data[i,4],data[i+1,4],data[i+1,5],data[i,5]))
```

```
      if(gene=="cytb"){
         polygon(x,y,col="plum3", border="plum3")
       }else{
        polygon(x,y,col="darkolivegreen3", border="darkolivegreen3")
       }
     }

#median value as a dashed line
points(log10(data[,3])~data[,1],type="l", lty=2, lwd=2)
#edge the HPD interval by plotting the upper and lower 95% HPD
points(log10(data[,4])~data[,1],type="l")
points(log10(data[,5])~data[,1],type="l")


title(main = plot_title)
```



Appendix C

Supplementary Information for Chapter 4

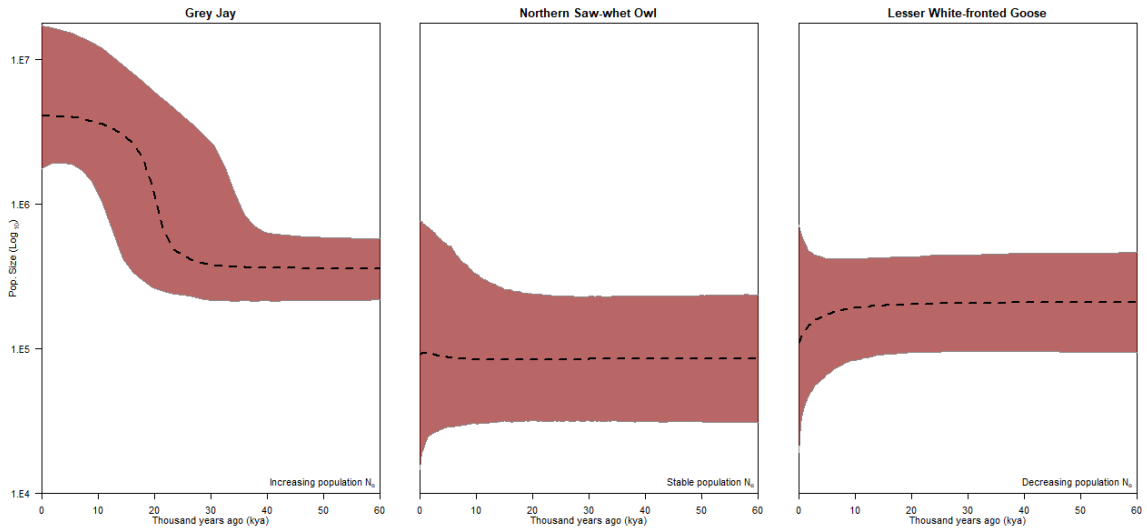


Fig. C.1 Example Bayesian skyline plots (BSPs) for the three categories of population trajectories; increasing effective population size (N_e), stable N_e, and decreasing N_e. Each population history is inferred from mitochondrial DNA data originally downloaded from GenBank. Dotted line is the median estimate of Ne and the edge of the coloured polygon show the boundary of the 95% highest posterior density (HPD) intervals. The x-axis represents time from the present in thousands of years. All plots are on the same scale.

Species name	Gene	bp	n	Hap no.	Dropped
Aegithalos caudatus	cytb	864	102	19	Drop
	ND2	**1014**	**184**	**36**	
Anas platyrhynchos	cytb	508	69	7	Drop
	ND2	**950**	**44**	**23**	
Anser cygnoides	cytb	1045	31	8	
	ND2	**1039**	**12**	**5**	**Drop**
Cardinalis cardinalis	cytb	999	25	22	Drop
	ND2	**1032**	**165**	**81**	
Carpodacus erythrinus	cytb	664	15	12	Drop
	ND2	**1038**	**190**	**115**	
Certhia familiaris	cytb	512	24	13	Drop
	ND2	**1041**	**17**	**9**	
Cinclus cinclus	cytb	845	136	29	
	ND2	**968**	**108**	**22**	**Drop**
Corvus corax	cytb	875	64	22	
	ND2	**1041**	**16**	**13**	**Drop**
Erithacus rubecula	cytb	878	92	30	Drop
	ND2	**1008**	**84**	**28**	
Ficedula parva	cytb	998	21	10	Drop
	ND2	**1040**	**93**	**7**	
Fringilla coelebs	cytb	520	127	31	
	ND2	**1014**	**77**	**33**	**Drop**
Geothlypis trichas	cytb	894	16	8	Drop
	ND2	**979**	**18**	**10**	
Leucosticte tephrocotis	cytb	881	40	7	Drop
	ND2	**1041**	**89**	**19**	
Melanerpes aurifrons	cytb	568	13	8	Drop
	ND2	**1037**	**10**	**9**	
Merops apiaster	cytb	869	171	59	
	ND2	**909**	**174**	**53**	**Drop**
Parus monticolus	cytb	887	158	41	Drop
	ND2	**959**	**159**	**42**	
Picoides tridactylus	cytb	414	31	11	
	ND2	**332**	**30**	**5**	**Drop**
Pinicola enucleator	cytb	1141	24	17	Drop
	ND2	**1040**	**37**	**28**	
	ND2	1040	39	24	
Poecile montanus	**cytb**	**829**	**60**	**30**	**Drop**
	ND2	994	189	78	
Poecile palustris	**cytb**	**920**	**110**	**48**	**Drop**
	ND2	996	103	31	
Regulus regulus	**cytb**	**558**	**45**	**10**	**Drop**
	ND2	1029	82	20	
Sturnella magna	**cytb**	**997**	**45**	**23**	
	ND2	1030	38	21	Drop
Sylviparus modestus	**cytb**	**948**	**18**	**11**	
	ND2	1038	18	14	Drop
Toxostoma curvirostre	**cytb**	**398**	**83**	**25**	
	ND2	225	86	19	Drop
Turdus merula	**cytb**	**626**	**85**	**27**	
	ND2	1035	71	24	Drop

Table C.1 For 25 species in analysis sequence data was available for both cytb and ND2 but data for one gene was included. Table shows which of the duplicate dataset were dropped and why.

Scientific Name	GBIF citation	Obs. In thinned dataset	LGM range size	Present range size	Overlap between periods	BSP rel. diff. 60kya-5kya
Acrocephalus palustris	https://doi.org/10.15468/dl.0ud3il	455	1670000	7930000	0.14092	0.24589
Acrocephalus schoenobaenus	https://doi.org/10.15468/dl.zeskgi	602	6222500	13590000	0.34216	0.11873
Actitis hypoleucos	https://doi.org/10.15468/dl.x83j5z	958	10725000	24370000	0.33309	22.35937
Aegithalos caudatus	https://doi.org/10.15468/dl.r4egg6	933	8047500	16345000	0.27807	0.65332
Aegolius acadicus	https://doi.org/10.15468/dl.7ze007	508	6100000	7375000	0.25085	0.02210
Alectoris rufa	https://doi.org/10.15468/dl.s3mkbk	133	5965000	5752500	0.56410	0.19171
Anser erythropus	https://doi.org/10.15468/dl.ilsiji	55	8137500	3257500	0.47659	-0.17505
Apus apus	https://doi.org/10.15468/dl.4ky8gs	1331	14940000	2.30E+07	0.56109	0.27763
Aythya fuligula	https://doi.org/10.15468/dl.mys46e	596	5725000	18152500	0.21581	0.12981
Bubo bubo	https://doi.org/10.15468/dl.rryfpw	782	17152500	28475000	0.47340	-0.24643
Burhinus oedicnemus	https://doi.org/10.15468/dl.uhvq5i	363	7972500	13500000	0.44074	10.80686
Calidris alpina	https://doi.org/10.15468/dl.ejolcu	138	8671250	5273750	0.33963	16.91847
Calidris maritima	https://doi.org/10.15468/dl.ep6qbn	60	5651250	2565000	0.33758	0.25891
Campylorhynchus brunneicapillus	https://doi.org/10.15468/dl.na3wwa	281	2465000	3142500	0.57200	4.35962
Cardinalis cardinalis	https://doi.org/10.15468/dl.am21mz	805	3947500	8165000	0.37416	5.48780
Carduelis carduelis	https://doi.org/10.15468/dl.xmetea	1011	4060000	11690000	0.27224	0.04373
Carpodacus erythrinus	https://doi.org/10.15468/dl.ysbswo	629	17080000	22587500	0.45656	82.06242
Catharus guttatus	https://doi.org/10.15468/dl.m8o4tm	895	5392500	8832500	0.19615	-0.02413
Catharus minimus	https://doi.org/10.15468/dl.gn08om	198	990000	5042500	0.06991	0.97093
Catharus ustulatus	https://doi.org/10.15468/dl.ep6qbn	71	642500	1017500	0.32187	4.29293
Certhia brachydactyla	https://doi.org/10.15468/dl.ahhkhi	385	2562500	5060000	0.38291	0.80796
Certhia familiaris	https://doi.org/10.15468/dl.qzbf4c	670	9587500	14772500	0.36267	0.98465
Chloris chloris	https://doi.org/10.15468/dl.2ebvzt	996	8340000	12650000	0.52016	0.89842
Cinclus cinclus	https://doi.org/10.15468/dl.avyo81	539	16425000	12952500	0.73577	3.93598
Clangula hyemalis	https://doi.org/10.15468/dl.y6n3ef	123	9743750	6080000	0.41893	0.36300
Coccothraustes coccothraustes	https://doi.org/10.15468/dl.ypgfiu	618	3825000	12260000	0.15987	1.33907
Coccyzus americanus	https://doi.org/10.15468/dl.zrb4yd	722	2765000	6597500	0.30580	0.24166
Colaptes auratus	https://doi.org/10.15468/dl.v0vavm	1104	2825000	10772500	0.19146	-0.13099
Columba palumbus	https://doi.org/10.15468/dl.lb5hna	1005	6227500	12832500	0.36158	0.07835
Coturnicops noveboracensis	https://doi.org/10.15468/dl.rrpg69	123	2142500	5497500	0.00000	4.12956
Cyanistes caeruleus	https://doi.org/10.15468/dl.vuqhcr	881	4492500	10365000	0.32079	0.03027
Emberiza schoeniclus	https://doi.org/10.15468/dl.fmwstu	738	8002500	17365000	0.34063	15.71831
Empidonax alnorum	https://doi.org/10.15468/dl.oeu1qu	605	2887500	7252500	0.03240	11.86088
Eremophila alpestris	https://doi.org/10.15468/dl.aowfc0	394	14225000	13970000	0.55260	22.69301
Erithacus rubecula	https://doi.org/10.15468/dl.6mqxj7	860	5442500	10905000	0.36749	15.84810
Eugenes fulgens	https://doi.org/10.15468/dl.cscfrx	90	2980000	2640000	0.89489	0.82515
Ficedula zanthopygia	https://doi.org/10.15468/dl.q6feoj	51	4792500	5810000	0.51979	0.17242
Fringilla coelebs	https://doi.org/10.15468/dl.v5vilo	1082	6705000	14080000	0.32546	0.48435
Geothlypis trichas	https://doi.org/10.15468/dl.bcsrme	1466	6400000	12507500	0.41335	0.26330
Gyps fulvus	https://doi.org/10.15468/dl.yfrkak	197	8427500	10817500	0.61914	1.48611
Icterus galbula	https://doi.org/10.15468/dl.vehq2a	580	2150000	6010000	0.16722	2.11099
Junco hyemalis	https://doi.org/10.15468/dl.q7ze67	1015	6480000	10442500	0.20613	2.67990
Junco phaeonotus	https://doi.org/10.15468/dl.nthkjc	75	3277500	2865000	0.87522	0.13765
Lanius collurio	https://doi.org/10.15468/dl.n0rgdf	752	3315000	10860000	0.22928	2.75837
Leucosticte brandti	https://doi.org/10.15468/dl.ehjnrx	63	15687500	4155000	0.96330	0.70985
Leucosticte nemoricola	https://doi.org/10.15468/dl.w5sekb	83	19555000	6120000	0.92402	3.89641
Leucosticte tephrocotis	https://doi.org/10.15468/dl.4uexnp	119	2360000	4385000	0.22121	0.94800
Locustella certhiola	https://doi.org/10.15468/dl.vadic8	86	11237500	15380000	0.61671	9.83482
Locustella luscinioides	https://doi.org/10.15468/dl.cyqxqb	279	1547500	7700000	0.11981	0.51077
Melospiza melodia	https://doi.org/10.15468/dl.c4yxws	1121	6037500	10510000	0.28235	0.92995
Merops apiaster	https://doi.org/10.15468/dl.khblcb	646	5655000	11905000	0.43028	34.39585
Motacilla alba	https://doi.org/10.15468/dl.lh5cn8	1987	28212500	37655000	0.63471	12.67849
Muscicapa striata	https://doi.org/10.15468/dl.x6kqo6	1006	7165000	14752500	0.39180	2.29586

Table C.2 A.

Parus major	https://doi.org/10.15468/dl.aovmeb	1317	12430000	23197500	0.40802	7.71383
Parus monticolus	https://doi.org/10.15468/dl.sgsrni	104	9597500	5182500	0.90159	1.83178
Passer domesticus	https://doi.org/10.15468/dl.opczfm	1974	15825000	30295000	0.46617	4.00281
Passer hispaniolensis	https://doi.org/10.15468/dl.ytgpiu	289	5332500	8367500	0.55303	4.78748
Passer montanus	https://doi.org/10.15468/dl.msvsze	1757	25822500	33362500	0.57655	0.28177
Passerculus sandwichensis	https://doi.org/10.15468/dl.jdef4f	1118	8222500	14212500	0.22797	7.95980
Perisoreus canadensis	https://doi.org/10.15468/dl.mj4lda	657	4530000	7920000	0.06534	10.14992
Phylloscopus collybita	https://doi.org/10.15468/dl.wp7qsu	730	4720000	10127500	0.31054	0.01334
Phylloscopus pulcher	https://doi.org/10.15468/dl.spin6q	65	7667500	4697500	0.93827	34.90976
Picoides tridactylus	https://doi.org/10.15468/dl.w46mgx	115	5163750	11461250	0.09532	3.41076
Pinicola enucleator	https://doi.org/10.15468/dl.olse1n	498	5817500	10152500	0.11786	5.06564
Piranga rubra	https://doi.org/10.15468/dl.yvkij6	500	1580000	4565000	0.26396	2.36316
Plectrophenax nivalis	https://doi.org/10.15468/dl.dm2uil	92	10626250	4801250	0.47975	0.79640
Pluvialis squatarola	https://doi.org/10.15468/dl.biwyki	24	8447500	3886250	0.48948	0.27167
Poecile carolinensis	https://doi.org/10.15468/dl.4msgkm	390	892500	3067500	0.19071	1.94570
Poecile gambeli	https://doi.org/10.15468/dl.8syzrh	327	3755000	4155000	0.49398	14.92504
Poecile montanus	https://doi.org/10.15468/dl.xuahrs	847	9445000	20372500	0.27329	40.81041
Poecile palustris	https://doi.org/10.15468/dl.yh3srx	556	6357500	10462500	0.28363	2.16636
Polioptila melanura	https://doi.org/10.15468/dl.z5qnkw	148	1747500	2665000	0.44278	0.14624
Prunella modularis	https://doi.org/10.15468/dl.28460g	605	7285000	9722500	0.51144	0.22464
Pyrrhocorax pyrrhocorax	https://doi.org/10.15468/dl.gwolxs	423	26367500	15980000	0.94337	13.95474
Regulus regulus	https://doi.org/10.15468/dl.znctre	710	10957500	12595000	0.42676	2.57112
Riparia riparia	https://doi.org/10.15468/dl.2ears0	848	6515000	16550000	0.25185	73.20941
Scolopax rusticola	https://doi.org/10.15468/dl.vtznsn	574	5032500	11900000	0.18676	56.59060
Sitta europaea	https://doi.org/10.15468/dl.rxqypw	879	8410000	20317500	0.30528	6.12661
Sitta pygmaea	https://doi.org/10.15468/dl.gbn9d4	229	4770000	4842500	0.61797	0.34977
Sphyrapicus ruber	https://doi.org/10.15468/dl.vhrzay	112	1480000	2025000	0.41728	0.88586
Spinus spinus	https://doi.org/10.15468/dl.2ejl2b	529	5982500	9770000	0.37615	18.58897
Streptopelia turtur	https://doi.org/10.15468/dl.qilxdi	879	7437500	16057500	0.41071	17.80239
Sturnella magna	https://doi.org/10.15468/dl.xkvo0e	743	3895000	6767500	0.41263	1.39758
Sturnella neglecta	https://doi.org/10.15468/dl.u38qjg	797	4947500	7625000	0.44393	0.83898
Sturnus vulgaris	https://doi.org/10.15468/dl.ftm9db	1027	7537500	15597500	0.36961	1.14346
Sylviparus modestus	https://doi.org/10.15468/dl.ykeyem	36	7390000	5935000	0.80623	1.63878
Tachycineta bicolor	https://doi.org/10.15468/dl.o1vdjc	1139	6112500	11675000	0.24325	16.87843
Toxostoma curvirostre	https://doi.org/10.15468/dl.lfrhky	300	2582500	3160000	0.57991	3.19067
Toxostoma redivivum	https://doi.org/10.15468/dl.7zljty	37	2902500	2017500	0.78067	3.75332
Troglodytes troglodytes	https://doi.org/10.15468/dl.fm6eql	931	12690000	14002500	0.56418	0.79133
Turdus philomelos	https://doi.org/10.15468/dl.3wnjst	792	6352500	13032500	0.34721	3.42735
Turdus torquatus	https://doi.org/10.15468/dl.bjwqxs	189	8650000	8570000	0.66365	0.58938
Vermivora cyanoptera	https://doi.org/10.15468/dl.8oy3u2	222	1052500	2237500	0.03352	0.09148
Vireo atricapilla	https://doi.org/10.15468/dl.meootf	48	2467500	3645000	0.43827	0.54512
Zonotrichia albicollis	https://doi.org/10.15468/dl.xjdy6v	503	2860000	6987500	0.00823	0.15614
Zonotrichia atricapilla	https://doi.org/10.15468/dl.io3kom	130	757500	2485000	0.10664	0.57114

Table C.3 B. Table lists species for which SDMs were created. Details on Global Biodiversity Information Facility (GBIF) DOIs, occurrence data available for each individual species, as well as reconstructed potential range size and associated relative change in population size from BSPs

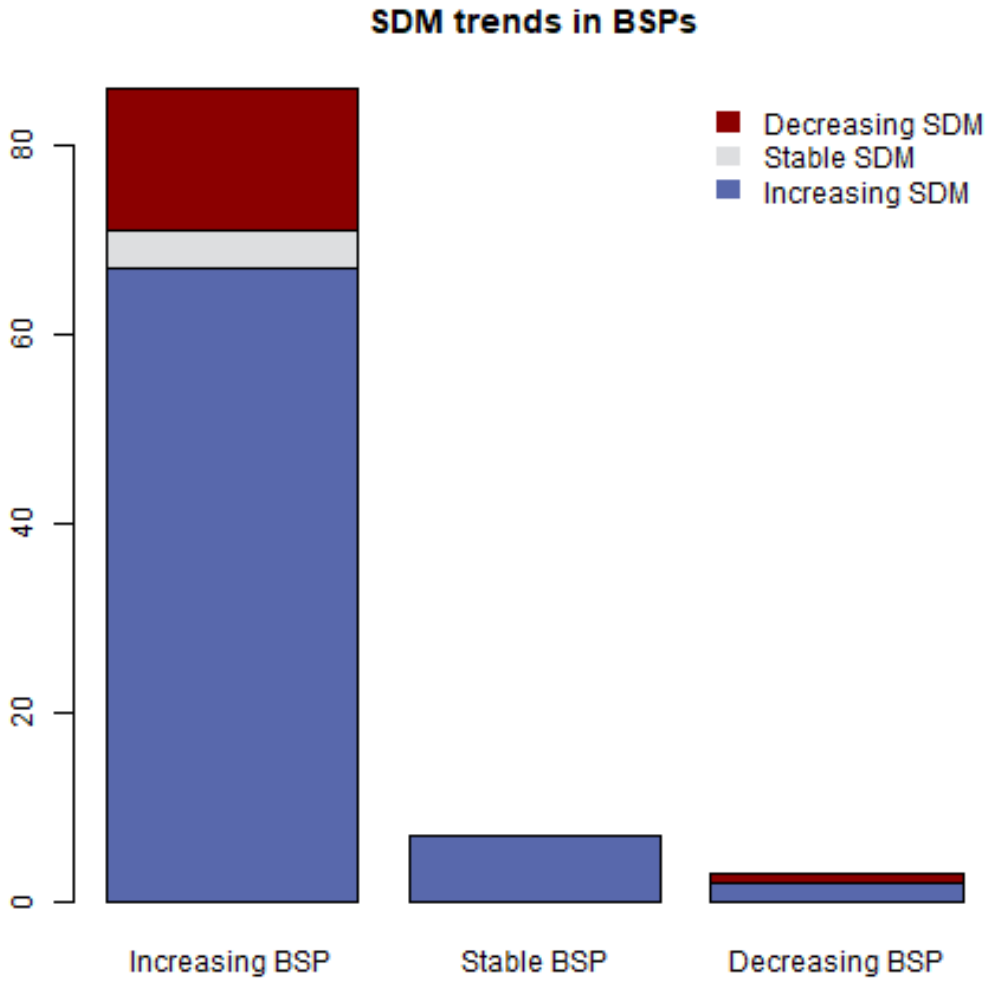


Fig. C.2 Barplot of BSP trends coloured by proportions of plots that have an increasing, stable or decreasing SDM trend. It was not possible to construct SDMs for all species, therefore $n = 96$.

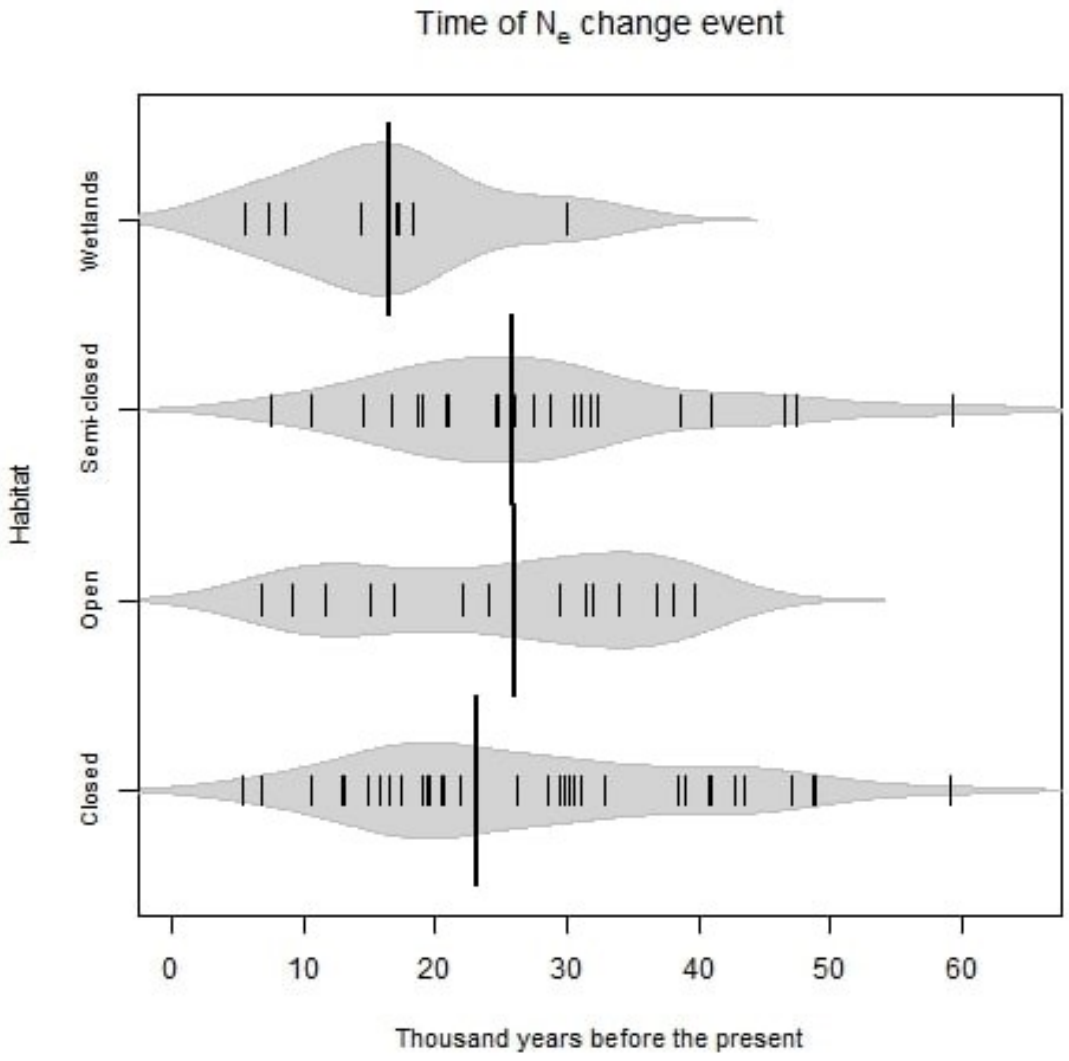


Fig. C.3 Beanplot showing time of dominant population change event for species from each habitat type, excluding populations with an overall change in N_e less than 10%. Kernels represent density, each small line the time of an individual population's size change event (increase or decrease). Thick black line is median time for change in the habitat. Bin sizes for both plots are; Closed ($n = 37$), Open ($n = 17$), Semi-closed ($n = 24$), Wetlands ($n = 12$).

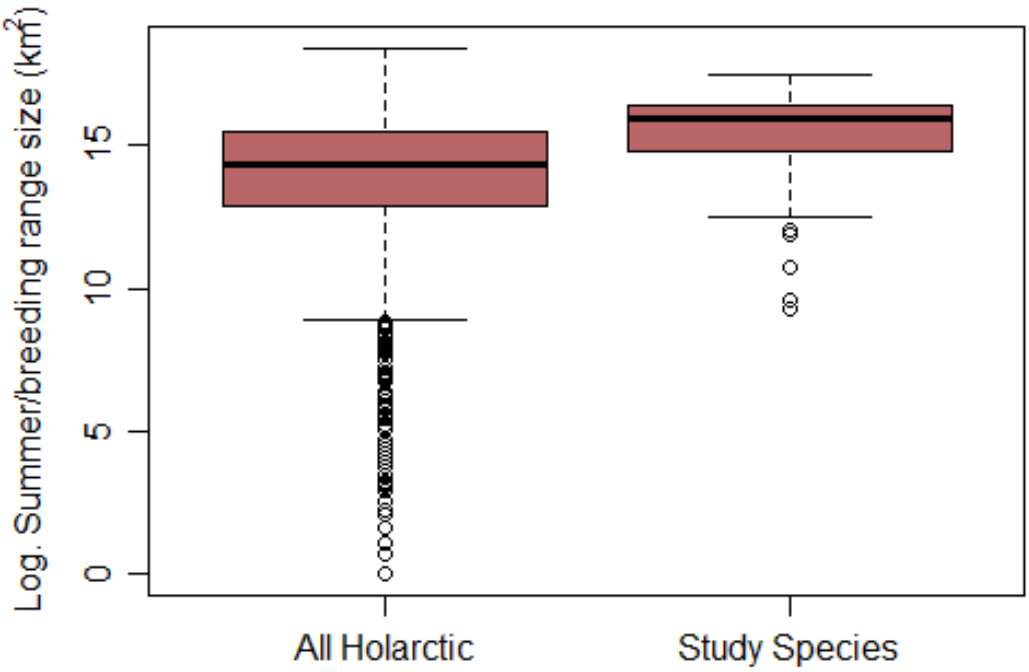


Fig. C.4 Box plot of Extent of Occurrence (EOO) for bird species with mean range latitude $\geq 20°$N, box one, and EOO for the 102 bird species for which we were able to construct BSP profiles, box two.

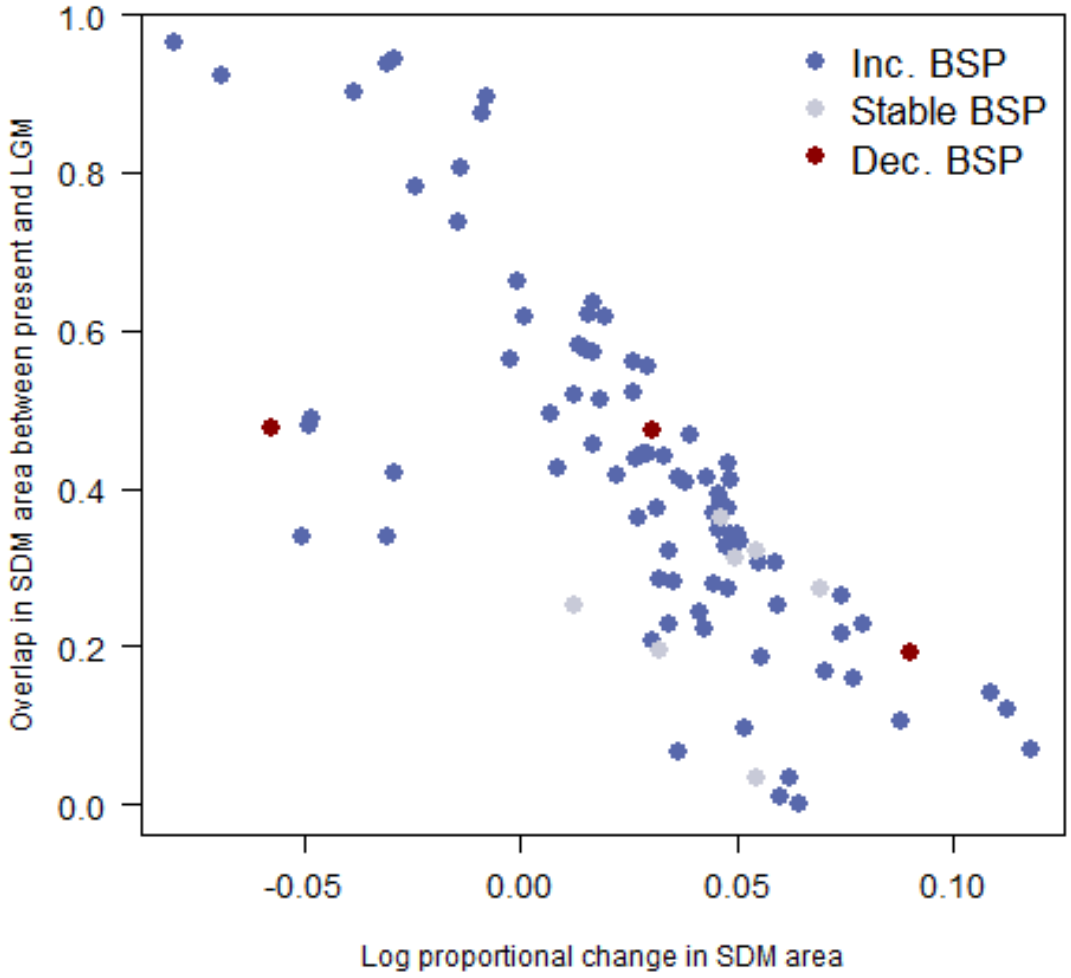


Fig. C.5 Scatter plot of change in overall SDM area and the proportion of each SDM in present that was also suitable for that species at the LGM (21 kya). Points are coloured by trend in N_e from BSP.

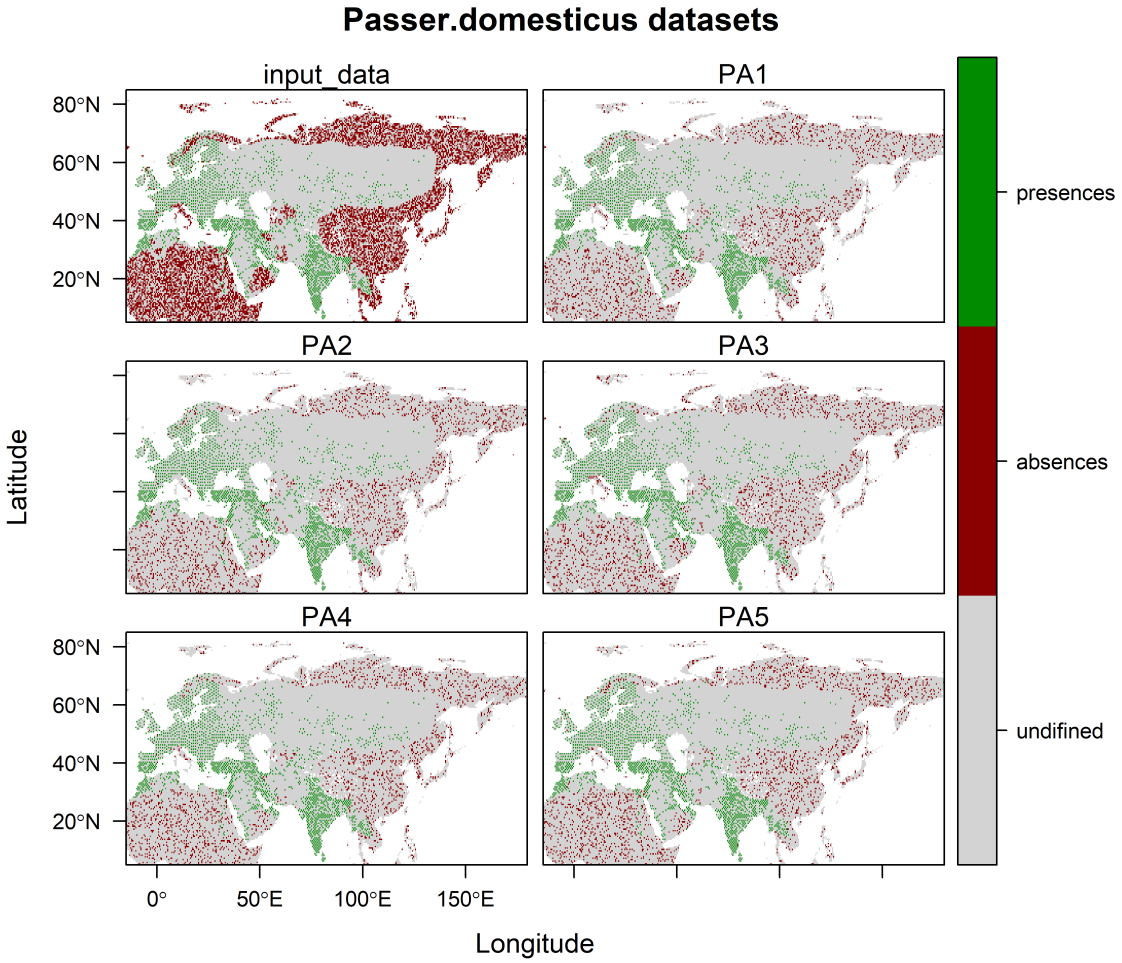


Fig. C.6 Example dataset for Species Distribution Model fitting using the BirdLife resident and breeding masks, Species shown is the house sparrow *Passer domesticus*. Panel one shows all data, PA1-5 show different sets of randomly sampled pseudoabsences.

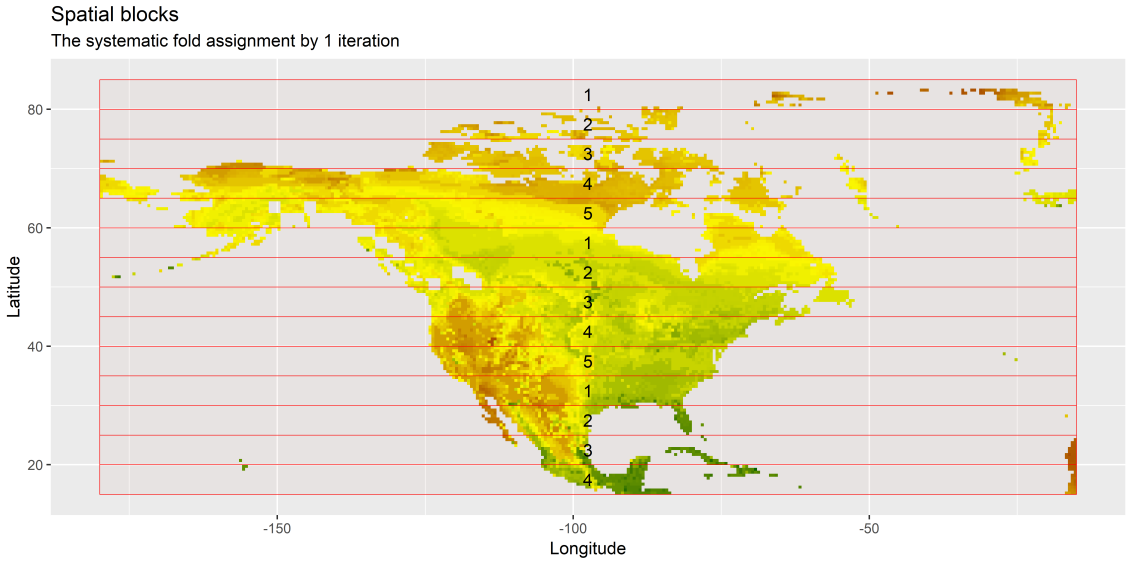


Fig. C.7 Spatial blocks based on latitudinal bands in North America, built with the R package *BlockCV* [203]

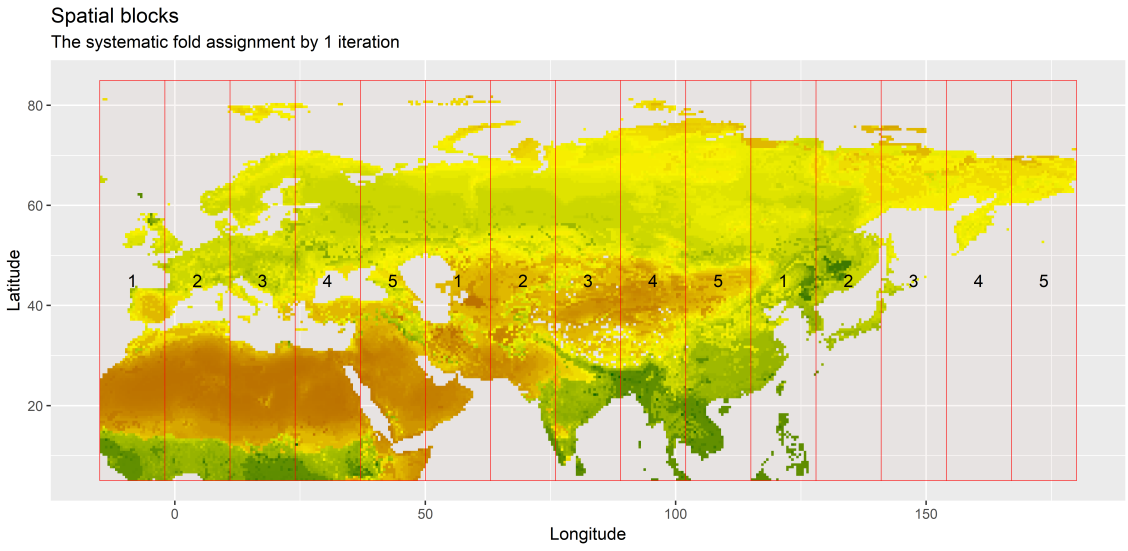


Fig. C.8 Spatial blocks based on longitudinal bands in Eurasia, built with the R package *BlockCV* [203]

```
#_____________________________________________
#
# SDM analyses of Holarctic birds
# Michela Leonardi, Department of Zoology, University of Cambridge.
# ml897@cam.ac.uk
#
# date: 2019-08-07
#
#_____________________________________________
#
setwd("C:/Users/miche/Desktop/SDM Birds/Holarctic birds 2019_08_07")

# libraries
library(rgbif)
library(rworldmap)
library(data.table)
library(PBSmapping)
library(rgeos)
library(maptools)
library(beanplot)
library(raster)
library(rgdal)
library(SDMTools)
library(RColorBrewer)
library(caret)
library(biomod2)
library(spThin)
library(blockCV)

# Function to modify coordinates
whereneарest <- function(val,matrix) {
  dist = abs(matrix-val)
  index = which.min(dist)
  return( index )
}

# Load world map
WorldMap <- getMap(resolution = "low")

# vector of variable names
vars <- c("npp","lai","BIO1","BIO4","BIO5","BIO6","BIO7","BIO8",
          "BIO9","BIO10","BIO11","BIO12","BIO13","BIO14","BIO15",
          "BIO16","BIO17","BIO18","BIO19")

```

```
45 # time for past projections (thousands of years ago)
46 kyr <- 21
47
48 # read species data (please move "Species.csv" from "Input_20190807"
49 # folder into working directory)
50 csv <- read.table("Species.csv",sep=",", header=TRUE)
51
52 for (i in 1:dim(csv)[1]){
53
54   # variables
55   sp <- as.character(csv$Species[i])
56   filename <- as.character(csv$GBIF.code[i])
57   biome <- as.character(csv$biome[i])
58   region <- as.character(csv$region[i])
59   outpath <- paste(sp, "/", region,"/", sep="")
60   dir.create(file.path(outpath),recursive=TRUE)
61
62   # Species name with "_" or "." instead of space
63   sp_name <- paste(unlist(strsplit(sp, " "))[1],
64                    unlist(strsplit(sp, " "))[2], sep="_")
65   sp.name <- paste(unlist(strsplit(sp, " "))[1],
66                    unlist(strsplit(sp, " "))[2], sep=".")
67
68   # define limit coordinates (W,E,S,N) and plot width
69   if (region=="NAmerica"){
70     coord <- c(-180,-15,8,90)
71     wtd <- 10
72     h <- 8
73   } else if (region=="Eurasia"){
74     coord <- c(-12,170,5,85)
75     wtd <- 13
76     h <- 6
77   } else {
78     stop('Region name is neiter "Eurasia" or "NAmerica",
79          please check spelling')
80   }
81
82   # define colors (filled and semi-transparent)
83   if (biome=="forest"){
84     color <- rgb(0.55,0.75,0.45,1)
85     Tcolor <- rgb(0.55,0.75,0.45,0.75)
86   } else if (biome=="grassland"){
87     color <- rgb(0,0.83,0,1)
88     Tcolor <- rgb(0,0.83,0,0.6)
```

```
89  } else if (biome=="shrubland"){
90    color <- rgb(0.65,0.5,0.2,1)
91    Tcolor <- rgb(0.65,0.5,0.2,0.45)
92  } else if (biome=="wetland"){
93    color <- rgb(0.4,0.7,0.8,1)
94    Tcolor <- rgb(0.4,0.7,0.8,0.75)
95  } else if (biome=="other"){
96    color <- rgb(0.5,0.7,0.7,1)
97    Tcolor <- rgb(0.5,0.7,0.7,0.75)
98  } else {
99    stop('Biome name is neither "forest", "grassland", "shrubland",
100   "wetland", or "other"; please check spelling')
101 }
102
103 # Environmental variables for the whole area, in the present
104 envdata <- read.table(paste("Input_20190807/", region,
105                             "/EnvirVar/EnvirVar_", region,"_0.txt",
106                             sep=""), header=TRUE, sep=" ")
107 colnames(envdata)[1:2] <- c("long","lat")
108
109 #Extract latitude and longitude
110 lon <- as.vector(unique(envdata$long))
111 lat <- as.vector(unique(envdata$lat))
112
113 #==================#
114 # Data preparation #
115 #==================#
116
117 # download GBIF file
118 occ_download_get(filename, overwrite=TRUE)
119
120 # unzip file
121 unzip(paste(filename,".zip",sep=""), overwrite = TRUE,
122       exdir = ".", unzip = "internal")
123
124 # read file
125 distrib <- fread(paste(filename,".csv",sep="")) #GBIF file
126
127 # filter by latitude and longitude
128 distrib <- distrib[distrib$decimalLongitude > min(lon) &
129              distrib$decimalLongitude < max(lon) &
130              distrib$decimalLatitude > min(lat) &
131              distrib$decimalLatitude < max(lat),]
132
```

```
# filter by coordinate uncertainty, and only keep relevant columns
distrib <- distrib[distrib$coordinateUncertaintyInMeters < 10000 |
            is.na(distrib$coordinateUncertaintyInMeter),
          c("gbifID","decimalLatitude","decimalLongitude")]

# modify lat. and long. to match the grid of the climatic reconstructions
distrib$decimalLatitude <- sapply(distrib$decimalLatitude, function(x)
  lat[wherenearest(x,lat)])
distrib$decimalLongitude <- sapply(as.numeric(distrib$decimalLongitude),
                                   function(x) lon[wherenearest(x,lon)])

# remove duplicates
distrib <- distrib[!duplicated(distrib[,c("decimalLatitude",
                                          "decimalLongitude")]),]

# read mask data
SpeciesMask<-readShapePoly(paste("Input_20190807/Masks/", sp_name, sep=""),
                           proj4string=CRS("+proj=longlat +datum=WGS84"))

sapply(slot(SpeciesMask, "polygons"), function(x) slot(x, "ID"))

# subset it for presence=1 or 2
SpeciesMask.sub<-SpeciesMask[SpeciesMask$PRESENCE<3,]
SpeciesMask.sub<-SpeciesMask[SpeciesMask$ORIGIN<3,]

# select the summer/resident component
SpeciesMask<-SpeciesMask.sub[SpeciesMask.sub$SEASONAL%in%c("1","2"),]

# create Polysets for summer and resident component:
# merging all polygons to create a single PID, needed for later operations
SpeciesMask<-SpatialPolygons2PolySet(SpeciesMask)
if (length(unique(SpeciesMask$PID))>1) {
  SpeciesMask2<-joinPolys(SpeciesMask,operation="UNION")
}
SpeciesMask<-PolySet2SpatialPolygons(SpeciesMask)

# Plot map with observations and mask
# plot(WorldMap, col="cornsilk", bg="lightblue1", border = "grey",
#     xlim = coord[1:2], ylim = coord[3:4], lwd=0.05)
# points(as.numeric(distrib$decimalLongitude),
#     as.numeric(distrib$decimalLatitude), pch="*", col=color)
# plot(SpeciesMask, add=TRUE)

# Subset data if within polygon
```

```
177 # define coordinates
178 xy <- distrib[,c(3,2)]
179
180 #transform into SpatialPointsDataFrame
181 df <- SpatialPointsDataFrame(coords=xy, data=distrib[,1],
182                              proj4string=SpeciesMask@proj4string)
183
184 # keep only points in shapefile
185 distrib <- df[!is.na(over(df,SpeciesMask)),]
186
187 # create file with obs (3 cols: long, lat, observation -1 as presence)
188 distrib <- cbind(distrib$decimalLongitude, distrib$decimalLatitude,
189                  rep(1, length(distrib$decimalLongitude)))
190
191 colnames(distrib) <- c("long","lat",sp_name)
192
193 # save file
194 write.table(distrib, paste(outpath,sp_name,"_distrib.txt", sep=""))
195
196 #======================#
197 # Ecological analyses #
198 #======================#
199
200 # Remove NA from environmental data
201 envdata <-  envdata[complete.cases(envdata), ]
202
203 # Extract variables for observations
204 obs <- merge(distrib[,c(1,2)], envdata)
205 #head(obs)
206
207 # merge data in a table, last col distinguish between baseline & obs
208 envdata[,"set"] <- "baseline"
209 obs[,"set"] <- "obs"
210 data <- rbind(envdata,obs)
211
212 # Variable distribution plot
213 png(paste(outpath, sp_name,"_variables.png", sep=""),
214     height=9, width=8,  units = 'in', res = 600)
215 par(mfrow = c(4, 5),
216     oma = c(0, 1, 5, 0),
217     mar = c(1, 1, 3, 1),
218     mgp = c(0.5, 0.5, 0))
219
220 # for each environmental variable
```

```
for (v in vars){

  # Beanplot distributions
  beanplot(data[data$set=="baseline",v],data[data$set=="obs",v],
           bw="nrd",side = "both", col=list("black",color),
           border = c("black", color), what=c(1,1,1,0),
           main = v ,xaxt = 'n')
}
plot.new()
legend("center", c("Baseline", "Species\noccurrences"),
       fill = c("black", color), border=NA, bty="n", cex=1.5)

title(main=paste(sp,"climatic variables", sep=" "),
      cex.main= 3, outer = TRUE, line = 2)
dev.off()

#====================#
# Cross-correlation #
#====================#

# function to plot, courtesy of Raquel A. Garcia, Stellenbosch University.
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

  test <- cor.test(x,y)
  # borrowed from printCoefmat
  Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
                   cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                   symbols = c("***", "**", "*", ".", " "))

  text(0.5, 0.5, txt, cex = cex * r)
  text(.8, .8, Signif, cex=cex, col=2)
}

# Pairwise correlation matrix
cormat <- cor(envdata[,vars])

# plot correlation for all variables
```

```
265 png(filename=paste(outpath, sp_name, "_correlation_all_vars.png",sep=""),
266     width=1200, height=900)
267 pairs(envdata[,vars], lower.panel=panel.smooth, upper.panel=panel.cor)
268 dev.off()
269
270 # define uncorrelated variables
271 uncor <- findCorrelation(cormat, cutoff = 0.7)
272
273 # plot correlation for chosen variables
274 png(filename=paste(outpath, sp_name,"_correlation_uncorr_vars.png",
275                    sep=""), width=1200, height=900)
276 pairs(envdata[,vars[-c(uncor)]], lower.panel=panel.smooth,
277       upper.panel=panel.cor)
278 dev.off()
279
280 #==========#
281 # Thinning #
282 #==========#
283
284 # thinning
285 t <- thin(loc.data=as.data.frame(distrib), lat.col="lat", long.col="long",
286           spec.col=sp_name, thin.par=70, reps=100,
287           locs.thinned.list.return=TRUE, write.files=TRUE, max.files=10,
288           out.dir=paste(outpath,"Thin/", sep=""), out.base=sp_name,
289           write.log.file=FALSE)
290
291 ThinDistrib <- read.table(paste(outpath, "Thin/", sp_name, "_thin1.csv",
292                                 sep=""),header=TRUE,  sep=",")
293 ThinDistrib <- ThinDistrib[,c(2,3,1)]
294
295 # plot thinning
296 png(paste(outpath, sp_name, "_thinned_dataset.png", sep=""),
297     height=6, width=wtd,  units = 'in', res = 600)
298 par(mfrow = c(1, 2),      # 1x2 layout
299     oma = c(0, 0, 5, 0), # rows at the outer bottom left top right margin
300     mar = c(3, 1, 3, 1), # row of text at ticks & to separate plots
301     mgp = c(2, 1, 0))     # axis label 2 rows out, tick labels at 1 row
302
303 plot(WorldMap, col="cornsilk", bg="lightblue1", border = "grey",
304      xlim = coord[1:2], ylim = coord[3:4], lwd=0.05,
305      main=paste("Dataset\nN =", dim(distrib)[1], sep=" "))
306 points(as.numeric(distrib[,"long"]),as.numeric(distrib[,"lat"]),
307        pch="*", col=color)
308
```

```
plot(WorldMap, col="cornsilk", bg="lightblue1", border = "grey",
     xlim = coord[1:2], ylim = coord[3:4], lwd=0.05,
     main=paste("Thinned dataset\nN =", dim(ThinDistrib)[1], sep=" "))
plot(SpeciesMask, add=TRUE, col=color, border=NA)
points(as.numeric(ThinDistrib$long), as.numeric(ThinDistrib$lat),
       pch="*", col="black")

title(main=paste(sp,"occurrences", sep=" "),cex.main= 3,
      outer = TRUE, line = 2)

dev.off()

#==================================#
# Formatting the data for biomod2 #
#==================================#

# Raster maps for climatic variables through time
basefile <- paste("Input_20190807/", region, "/", sep="")

# Color palette
c1 <- colorRampPalette(c("khaki1",color))(6)
c2 <- colorRampPalette(c(color,"black"))(8)
cols <- c(c1[2:5],c2[2:8])

# environmental variables
expl.var <- stack()

vars1 <- vars[-c(uncor)]
# create raster stack
for(v in 1:length(vars1)){
  r <- raster(paste(basefile, vars1[v], "/", region, "_", vars1[v], "_0",
                    ".grd", sep=""), RAT = FALSE)
  expl.var <- stack( expl.var, r)
}

#==========================#
# Pseudo-absence selection #
#==========================#

# background data
bg <- data.frame(cbind(envdata[,1:2], species=rep(0,dim(envdata)[1])))
colnames(bg)[3] <- sp.name

# define coordinates
```

```
xy <- bg[,1:2]

#transform into SpatialPointsDataFrame
df <- SpatialPointsDataFrame(coords=xy, data=bg,
                             proj4string=SpeciesMask@proj4string)

# remove points within the mask
bg.clean <- df[is.na(over(df,SpeciesMask)),]
bg.clean <- data.frame(bg.clean[,1:3])

# # plot
# plot(WorldMap, col="cornsilk", bg="lightblue1", lwd=0.05,
#      border = "grey", xlim = c(min(lon), max(lon)), ylim = c(min(lat),
#      max(lat)))
# points(as.numeric(bg.clean[,"long"]), as.numeric(bg.clean[,"lat"]),
#       pch="*", col="red")

# number of absences and presences
pres <- dim(ThinDistrib)[1]
abs <- dim(ThinDistrib)[1]

# Absences table
abs.table <- data.frame(matrix("FALSE",abs*5,7), stringsAsFactors=FALSE)
colnames(abs.table) <- c("long","lat",paste("RUN",c(1:5),sep=""))

# variable to define the first line to be written
start <- 1

for (k in 1:5){
  index <- sample(1:dim(bg.clean)[1], abs)
  abs.table[seq(start,abs*k),1:2] <- bg.clean[index,1:2]
  abs.table[seq(start,abs*k),2+k] <- rep("TRUE",abs)
  start <- (abs*k)+1
}

#====================#
# Species input file #
#====================#

#presences + background
resp.var <- as.numeric(c(rep(1, pres),rep(0,abs*5)))

resp.xy <- rbind(ThinDistrib[,c("long","lat")],
                 abs.table[,c("long","lat")])
```

```
397
398 # add presences to PA.table
399 pres.table <- data.frame(cbind(ThinDistrib[,c("long","lat")]),
400                          matrix("TRUE",pres,5))
401 colnames(pres.table) <- c("long","lat",paste("RUN",c(1:5),sep=""))
402
403 # merge tables for presences and pseudoabsences
404 PA.table <- rbind(pres.table[,-c(1,2)], abs.table[,-c(1,2)])
405 PA.table[] <- lapply(PA.table, as.logical)
406
407 #========================#
408 # Input data for Biomod2 #
409 #========================#
410
411 biomodData <- BIOMOD_FormatingData(resp.var=resp.var, # species dist
412                                    expl.var=expl.var, # env vars
413                                    resp.xy=resp.xy,   # coords of species
414                                    resp.name=sp.name, # species col
415                                    # PA is pseudo absences
416                                    PA.strategy="user.defined",
417                                    PA.table=PA.table,
418                                    na.rm=TRUE) # do not consider NA
419
420 png(paste(outpath, sp_name, "_pseudoabsences_datasets.png", sep=""),
421     width=7, height=7, units='in', res=600)
422 plot(biomodData)
423 dev.off()
424
425 # To store the resulting object
426 save(biomodData, file=paste(outpath, sp_name, "_BiomodData", sep=""))
427
428 #========================#
429 # Block cross-validation #
430 #========================#
431
432 # presences and absences taken from biomodData (with NA removed)
433 resp <- cbind(biomodData@coord,sp=biomodData@data.species)
434 resp[] <- lapply(resp, as.numeric)
435 colnames(resp)[3] <- sp
436
437 # transform into SpatialPointsDataFrame
438 resp <- SpatialPointsDataFrame(resp[,c("long", "lat")], resp,
439                                proj4string=crs(expl.var))
440
```

```
441 # create spatial blocks and saves plot
442 png(paste(outpath,sp_name,"_spatial_blocks.png",sep=""),
443     height=6, width=10,  units='in', res=600)
444
445 # If NAmerica horizontal blocks, if Eurasia vertical blocks
446 if (region=="NAmerica"){
447   sb <- spatialBlock(speciesData = resp,
448                      species = sp,
449                      rasterLayer = expl.var,
450                      rows = 15,
451                      k = 5,
452                      selection = "systematic",
453                      biomod2Format = TRUE)
454
455 } else if (region=="Eurasia") {
456   sb <- spatialBlock(speciesData = resp,
457                      species = sp,
458                      rasterLayer = expl.var,
459                      cols = 15,
460                      k = 5,
461                      selection = "systematic",
462                      biomod2Format = TRUE)
463 }
464 dev.off()
465
466 # !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! #
467 # WARNING:                                                                  #
468 # it is necessary to check if all the runs contain presences #
469 # (>10) and absences. If not the table must be subset within #
470 # the BIOMOD_Modeling function (see next warning, line 491)  #
471 # !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! #
472
473 DataSplitTable <- sb$biomodTable
474
475 # add coordinates to the table
476 dst.xy <- cbind(resp$long,resp$lat,DataSplitTable)
477 colnames(dst.xy)[1:2] <- c("x", "y")
478
479 # plot
480 png(paste(outpath,sp_name,"_blocks_CV.png",sep=""), width=7, height=h,
481     units='in', res=600)
482 par(mfrow=c(3, 2),
483     oma=c(0, 0, 5, 0),
484     mar=c(1, 1, 1, 1),
```

```
485       mgp=c(2, 1, 0))
486
487   for (k in c(1:5)){
488
489     col <- DataSplitTable[,paste("RUN",k,sep="")]
490     col[col=="FALSE"] <- color
491     col[col=="TRUE"] <- "black"
492
493     plot(WorldMap, col="cornsilk", bg="lightblue1", lwd=0.05,
494          border = "grey", xlim = c(min(lon), max(lon)),
495          ylim = c(min(lat), max(lat)), main=paste("RUN",k,sep=""))
496     points(as.numeric(dst.xy[,"x"]), as.numeric(dst.xy[,"y"]),
497            pch="*", col=col)
498   }
499
500   # title
501   title(main=paste(sp," spatial blocks cross-validation",sep=""),
502         cex.main= 2, outer=TRUE, line=3)
503
504   dev.off()
505
506   #========================#
507   # Calibrating the models #
508   #========================#
509
510   # Model parameters
511   ModOptions <- BIOMOD_ModelingOptions()
512
513   mods <- c("GLM","GBM","GAM","RF") # GBM = boosted regression tree
514
515   ModelOut <- BIOMOD_Modeling(biomodData,              # input data
516                               models=mods,             # algorithms
517                               models.options=ModOptions, # options
518   # DataSplitTable WARNING: only consider columns (=splits) including
519   # presences (>10) and absences
520                               DataSplitTable=DataSplitTable,
521                               # DataSplitTable issued by blockCV
522                               models.eval.meth=c("TSS"),  # method for eval
523                               SaveObj=T,
524                               modeling.id=paste(sp.name,"_modelOut", sep=""),
525                               do.full.models=FALSE,
526                               rescal.all.models=T)
527   # model evaluations
528   ModelEvaluation <- get_evaluations(ModelOut)
```

```
529 ModelEvaluation["TSS","Testing.data",,,]
530
531
532 # TSS scores
533 write.table(ModelEvaluation["TSS","Testing.data",,,],
534             paste(outpath, sp_name, "_models_eval.txt", sep=""))
535
536 #====================================#
537 # Projecting to the entire study area #
538 #====================================#
539
540 new.env <- expl.var
541 Projection <- BIOMOD_Projection(modeling.output = ModelOut, # model
542                           selected.models = "all", # models to select
543                           new.env =  new.env, # env data to project to
544                           proj.name = region,
545                           binary.meth = c("TSS"),   # binary transformation
546                           filtered.meth = c("TSS"), # set values below zero
547                           build.clamping.mask = T,  # out-of-calibration
548                           compress = "xz")
549
550 # # make some plots sub-selected by str.grep argument
551 # png(paste(outpath,sp_name,"_proj_RF_RUN1.png", sep=""),
552 #     width=9, height=12, units='in', res=600)
553 # plot(Projection, str.grep = 'RF')
554 # dev.off()
555
556 #====================#
557 # Ensemble modelling #
558 #====================#
559
560 Ensemble <- BIOMOD_EnsembleModeling(modeling.output=ModelOut,
561                                     chosen.models="all",     # models used
562                                     em.by="all",             # by algorithm
563                                     eval.metric=c("TSS"),    # evaluation metric
564                                     eval.metric.quality.threshold=c(0.7),
565                                     models.eval.meth=c("TSS"),# evaluation meth
566                                     prob.mean=T,
567                                     prob.median=T,
568                                     committee.averaging=T,
569                                     prob.mean.weight=T,
570                                     prob.mean.weight.decay="proportional")
571
572 #=========================#
```

```
573 # Validation of the model #
574 #=========================#
575
576 # evaluation scores for the ensemble model
577 EnsembleEvaluation <- get_evaluations(Ensemble)
578 write.table(EnsembleEvaluation,
579             file=paste(outpath, sp_name,"_Ensemble_eval.txt", sep=""))
580
581 # change dimension names to make them more easily accessible
582 names(EnsembleEvaluation) <- c("mean_TSS","median_TSS",
583                                "ca_TSS","wmean_TSS")
584
585 # Create a table with evaluation scores, sensitivity and specificity
586 out<- rbind(c(EnsembleEvaluation$mean_TSS[,1],
587               EnsembleEvaluation$median_TSS[,1],
588               EnsembleEvaluation$ca_TSS[,1],
589               EnsembleEvaluation$wmean_TSS[,1]),
590               c(EnsembleEvaluation$mean_TSS[,3],
591               EnsembleEvaluation$median_TSS[,3],
592               EnsembleEvaluation$ca_TSS[,3],
593               EnsembleEvaluation$wmean_TSS[,3])/100,
594               c(EnsembleEvaluation$mean_TSS[,4],
595               EnsembleEvaluation$median_TSS[,4],
596               EnsembleEvaluation$ca_TSS[,4],
597               EnsembleEvaluation$wmean_TSS[,4])/100)
598
599 # index for the maximum TSS
600 maxTSS <- which.max(out[1,])
601
602 # Plot barplots of evaluation
603 colnames(out) <- c("Mean","Median","Committee\naverage","Weighted\nmean")
604 rownames(out) <- c("Evaluation score", "Sensitivity","Specificity")
605
606 png(paste(outpath, sp_name,"_evaluation_scores_ensemble.png", sep=""),
607     width = 6, height = 4.5, units = 'in', res = 600)
608 layout(matrix(c(1,2), ncol=1, byrow=TRUE), heights=c(3.8,0.7))
609 par(oma = c(0, 0, 4, 0), #rows of text at outer bottom left top right marg
610     mar = c(2, 3, 3, 1)) #rows of text at ticks and to separate plots
611
612 barplot(t(out), beside=TRUE, col=cols[c(6,2,8,4)], border=NA, ylim=c(0,1))
613
614 par(mai=c(0,0,0,0))
615 plot.new()
616 legend("center", colnames(out), fill=cols[c(6,2,8,4)], border=NA, bty="n",
```

```
        ncol=4)

title(main=paste("Evaluation scores -", sp, sep=" "), cex.main=1.5,
      outer=TRUE, line=1)

dev.off()

#======================#
# Ensemble forecasting #
#======================#

EnsembleProjection <- BIOMOD_EnsembleForecasting(EM.output = Ensemble,
                                                 projection.output = Projection,
                                                 selected.models = "all",
                                                 binary.meth = c("TSS"),
                                                 filtered.meth = c("TSS"),
                                                 compress = T)

#list of files in directory
filelist <- list.files(paste(sp.name,"/proj_",region,"/", sep=""))
# selecting consensus projections present
enslist <- filelist[grep("ensemble", filelist)]
PresEnsemble <- stack(paste(sp.name,"/proj_",region,"/",
                            enslist[2], sep=""))

EnsembleMethods <- c("Mean", "Median", "Committee average",
                     "Weighted mean")

# plot without observation points
png(paste(outpath, sp_name,"_ensemble_projection.png", sep=""),
    width = 12, height = 10, units = 'in', res = 600)
par(mfrow = c(2, 2),
    oma = c(0, 0, 4, 0),
    mar = c(3, 3, 2, 5),
    mgp = c(2, 1, 0))

for (x in 1:4) {
  #plot selected rasters (ensemble models)
  plot(raster(PresEnsemble,x),
       main=EnsembleMethods[x])
}
title(main=paste("Ensemble projections -", sp, sep=" "), cex.main= 1.5,
      outer = TRUE, line = 1)
```

```
dev.off()

# plot with observation points
png(paste(outpath, sp_name,"_ensemble_projection_points.png", sep=""),
    width = 12, height = 10, units = 'in', res = 600)
par(mfrow = c(2, 2),
    oma = c(0, 0, 4, 0),
    mar = c(3, 3, 2, 5),
    mgp = c(2, 1, 0))

for (x in 1:4) {
  #plot selected rasters (ensemble models)
  plot(raster(PresEnsemble,x),
       main=EnsembleMethods[x])
  points(ThinDistrib[,"long"], ThinDistrib[,"lat"], pch="*")
}

title(main=paste("Ensemble projections -", sp, sep=" "), cex.main= 1.5,
      outer = TRUE, line = 1)

dev.off()

#==============================#
# Projection backwards in time #
#==============================#

# environmental variables
past.env <- stack()

# create raster stack
for(v in 1:length(vars1)){
  r <- raster(paste(basefile, vars1[v], "/", region, "_",vars1[v],"_",
                    kyr, ".grd", sep=""), RAT = FALSE)
  past.env <- stack(past.env, r)
}

# Project model into the past
PastProjection <- BIOMOD_Projection(modeling.output = ModelOut,
                                    selected.models = "all",
                                    new.env =  past.env,
                                    proj.name = paste(kyr,region, sep="_"),
                                    binary.meth = "TSS",
                                    filtered.meth = "TSS",
                                    build.clamping.mask = T,
```

```
                                    compress = "xz")

# Project ensemble to the past
PastEnsemble <- BIOMOD_EnsembleForecasting(EM.output = Ensemble,
                                           projection.output = PastProjection,
                                           selected.models = "all",
                                           binary.meth = c("TSS"),
                                           filtered.meth = c("TSS"),
                                           compress = T)


filelistP <- list.files(paste(sp.name,"/proj_",kyr,"_", region,
                              "/", sep=""))
enslist <- filelistP[grep("ensemble", filelistP)]
PastEnsemble <- stack(paste(sp.name, "/proj_", kyr, "_",
                            region, "/", enslist[2], sep=""))

# plot all rasters
png(paste(outpath, sp_name,"_",kyr,"_all_ensemble_projections.png",
          sep=""), width = 12, height = 9, units = 'in', res = 600)

op<-par(mfrow = c(2, 2),
        oma = c(0, 0, 4, 1),
        mar = c(3, 3, 2, 5),
        mgp = c(2, 1, 0))

for (x in 1:4) {
  #plot selected rasters (ensemble models)
  plot(raster(PastEnsemble,x),
       main=EnsembleMethods[x])
}
title(main=paste("Ensemble projections LGM -", sp, sep=" "),
      cex.main= 1.5, outer = TRUE, line = 1)

dev.off()

#create LGM directory
dir.create(paste(outpath,"LGM",sep=""), showWarnings=FALSE)

# one plot for each statistic
for (x in 1:4) {
  png(paste(outpath, "LGM/",sp_name,"_",kyr,"_",EnsembleMethods[x],
            "_ensemble_projection.png", sep=""),
      width = 12, height = 9, units = 'in', res = 600)
```

```
749     plot(raster(PastEnsemble,x),
750          main=paste(sp, kyr, "kya", EnsembleMethods[x],
751                     "ensemble projection", sep=" "))
752     dev.off()
753   }
754
755   # index of the layers to use
756
757   for (j in 1:4){
758     # present
759     PresRaster <- raster(PresEnsemble,j)
760     writeRaster(PresRaster, paste(outpath,"LGM/", sp_name, "_",
761                                   EnsembleMethods[j], "_pres", sep=""),
762                 overwrite=TRUE)
763
764     # past
765     PastRaster <- raster(PastEnsemble,j)
766     writeRaster(PastRaster, paste(outpath,"LGM/", sp_name, "_",
767                                   EnsembleMethods[j], "_LGM", sep=""),
768                 overwrite=TRUE)
769
770     # difference between the two (present - past)
771     difference <- PresRaster-PastRaster
772     writeRaster(difference, paste(outpath, "LGM/", sp_name, "_",
773                                   EnsembleMethods[j], "_pres-past", sep=""),
774                 overwrite=TRUE)
775   }
776   # write on file
777   # clean dataset
778   csv[i,"Clean.dataset"] <- dim(distrib)[1]
779   # thinned dataset
780   csv[i,"Thinned.dataset"] <- dim(ThinDistrib)[1]
781
782   # Overwrite species file including results
783   write.csv(csv, file="Species.csv", quote = FALSE, row.names = FALSE)
784
785 }
```

Appendix D

Supplementary Information for Chapter 5

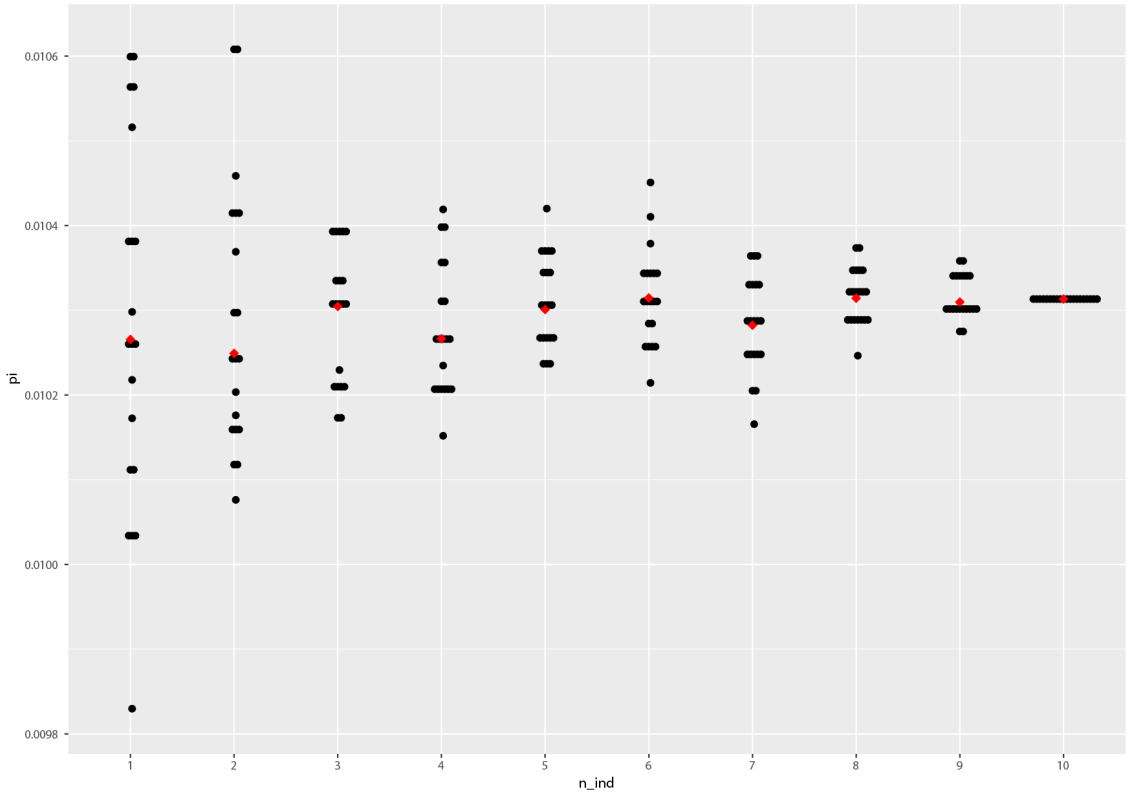


Fig. D.1 Plot of the pairwise π between two populations. Black dots represent the 20 replicates for each number of individuals, points in red show the median value.

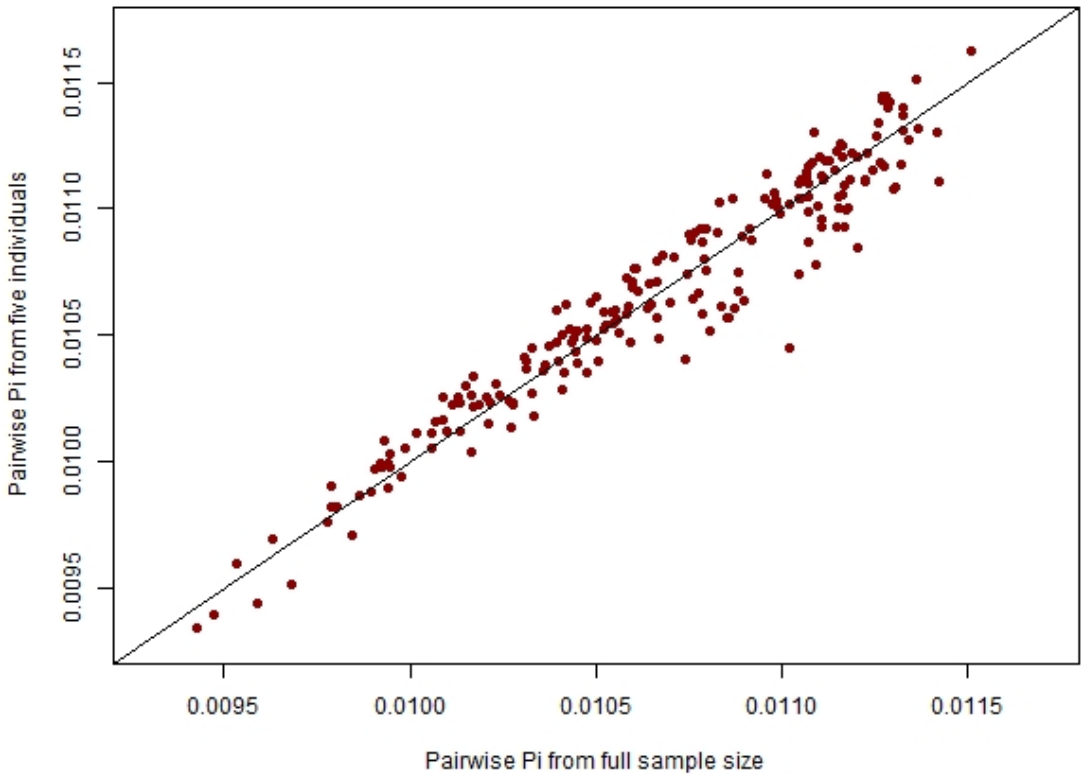


Fig. D.2 Relationship between pairwise π calculated from full sample sizes and pairwise π calculated from five individuals per population.

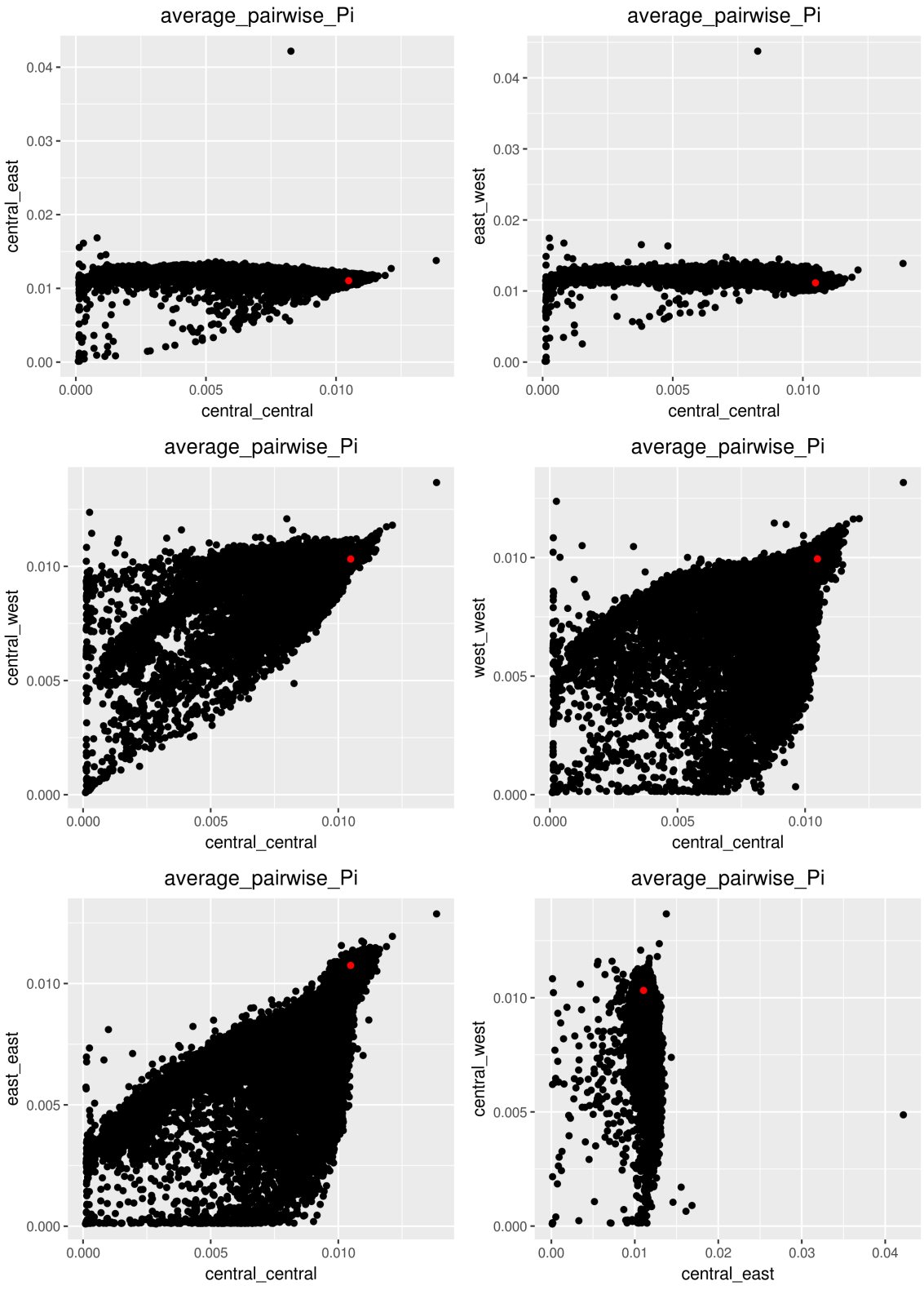

average_pairwise_Pi
central_east
central_central
average_pairwise_Pi
east_west
central_central
average_pairwise_Pi
central_west
central_central
average_pairwise_Pi
west_west
central_central
average_pairwise_Pi
east_east
central_central
average_pairwise_Pi
central_west
central_east

Fig. D.3 A.

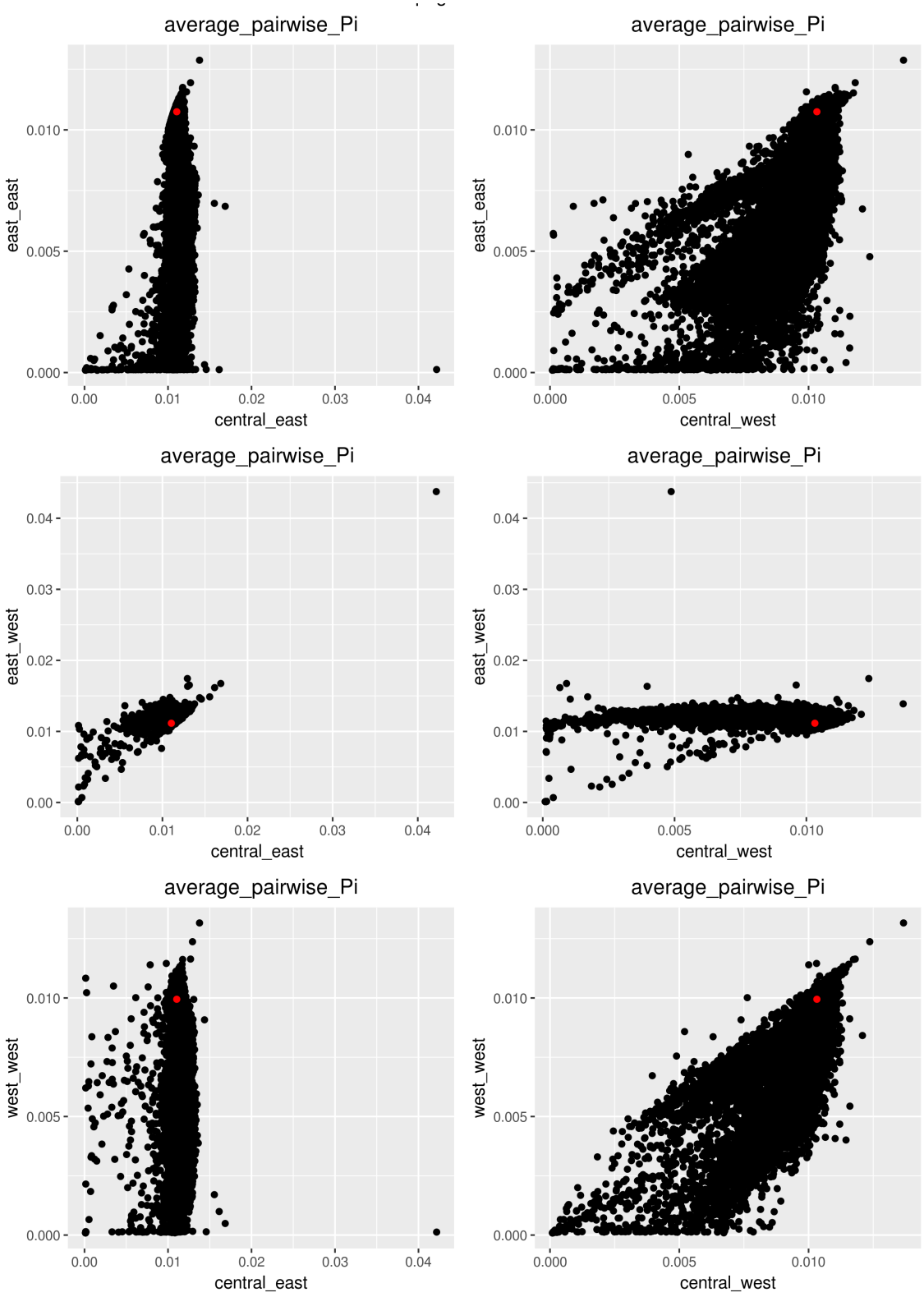

average_pairwise_Pi
east_east
central_east
average_pairwise_Pi
east_east
central_west
average_pairwise_Pi
east_west
central_east
average_pairwise_Pi
east_west
central_west
average_pairwise_Pi
west_west
central_east
average_pairwise_Pi
west_west
central_west

Fig. D.4 B.

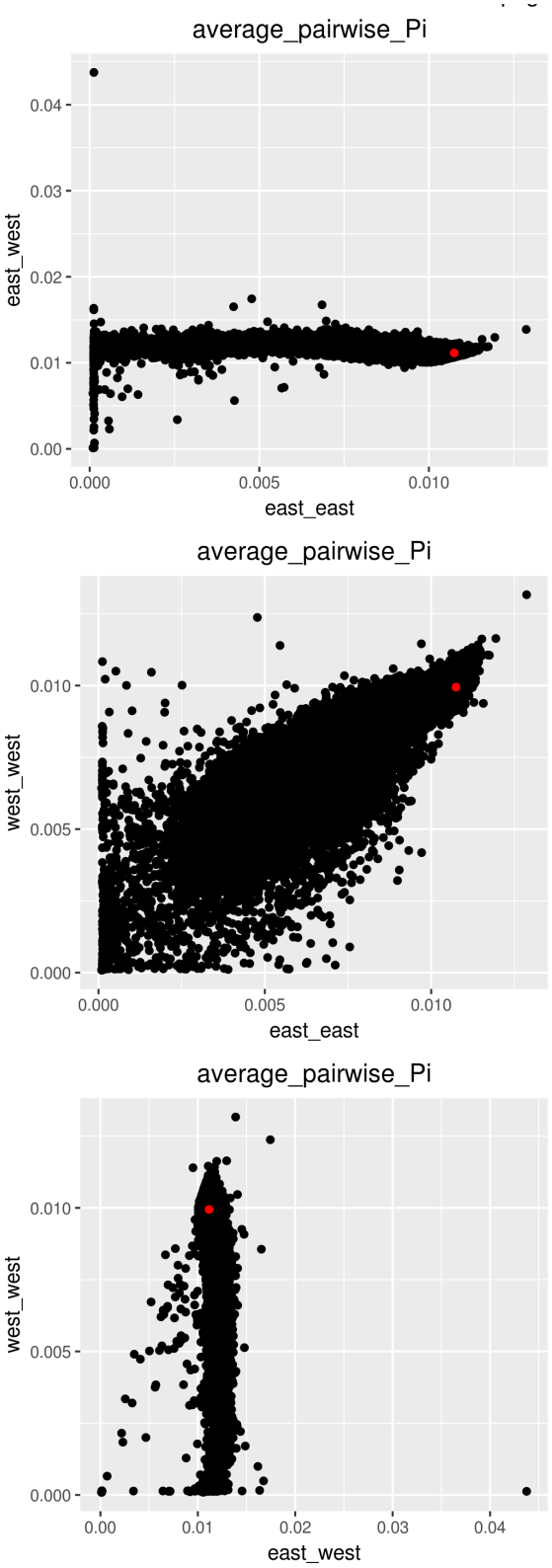


Fig. D.5 C. Pairwise plots of summary statistics distributions from a Monte-Carlo sweep. Observed values in red, simulated values in black.

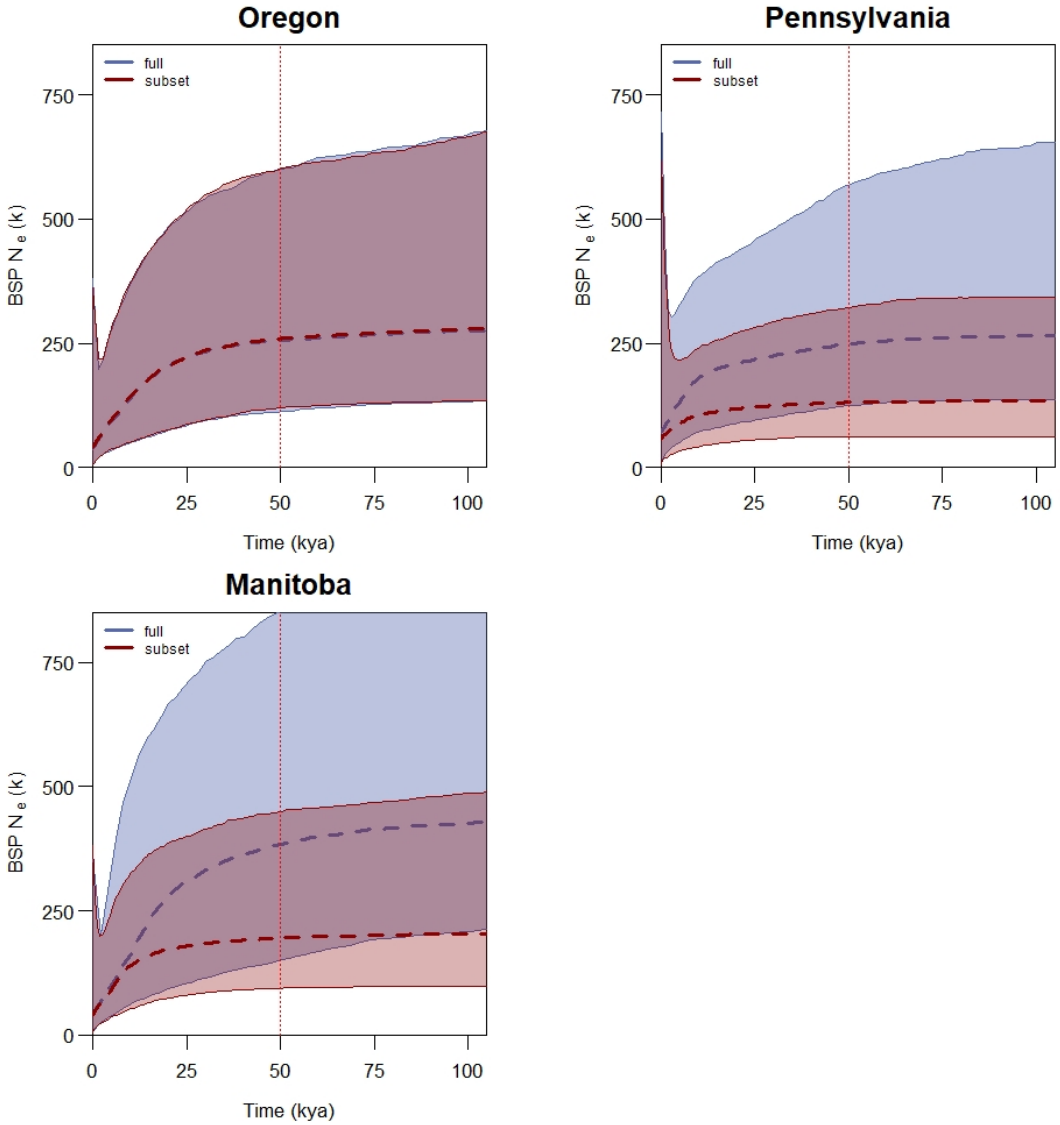


Fig. D.6 BSPs drawn for the three selected populations, mtDNA generated with Simulation 20 settings. Full dataset plotted in blue, subset dataset plotted in red, dashed line at center of each coloured polygon is the median value, edges of polygon represent the 95% Highest Posterior Density (HPD) interval. Dotted red line shows the start of the simualtion at 50 kya.

SDM analyses of Setophaga petechia

Michela Leonardi, Department of Zoology, University of Cambridge. ml897@cam.ac.uk

2019-05-20

Introduction

The following code has been used to analyse GBIF data from the American Yellow Warbler *Setophaga petechia* (former *Dendroica petechia*). The first chapter describes the code used to clean and format the species data from the *.csv format downloaded from the GBIF database for the analyses performed. The second chapter details a set of ecological analyses needed to better understand the effect of different climatic variables on the distribution of the species. Those latter are also preparatory for the species distribution modelling (third chapter) performed with the package biomod2.

Load and prepare the species data

Setting the working directory and variables, load a low resolution world map

```
library(rworldmap)
# Load world map
newmap <- getMap(resolution="low")

# vector of variable names
vars <-c("elevation","npp","lai","BIO1","BIO4","BIO5","BIO6","BIO7","BIO8","BIO9",
         "BIO10","BIO11","BIO12","BIO13","BIO14","BIO15","BIO16","BIO17","BIO18","BIO19")
```

A database with recorded presences of the species has been downloaded from GBIF (GBIF.org (19th November 2018) GBIF Occurrence Download https://doi.org/10.15468/dl.jfkwcg) and is available at this link.

Reading the file will give an error message that can be ignored.

```
db <- data.table::fread("0005643-181108115102211.csv") #GBIF file
```

The database contains a total of 1573147 observations. Some of them have one or more "issues" reported in the corresponding column. The following code compares the number and the distribution of observations without issues and with so-called "rounded coordinates".

```
# to check which kind of issues the data has
issue <- unique(db$issue)
#issue[1:4]

# remove all lines with reported issues and above the equator
db1 <- db[db$issue=="" & db$decimalLatitude > 0,]
# remove lines where uncertainty of coordintes is > 1000
db1b <- db1[db1$coordinateUncertaintyInMeters < 1000 |
              is.na(db1$coordinateUncertaintyInMeters),]

# keep lines with "coordinates rounded"
db2 <- db[db$issue%in%issue[c(1,2)] & db$decimalLatitude > 0,]
# remove lines where uncertainty of coordintes is > 1000
db2b <- db2[db2$coordinateUncertaintyInMeters < 1000 |
              is.na(db2$coordinateUncertaintyInMeters),]
```

Plot of a map comparing observations with and without rounded coordinates; they contain respectively 1526468 and 253745 observations. The extent of the region of interest (North America) is between -180°E,

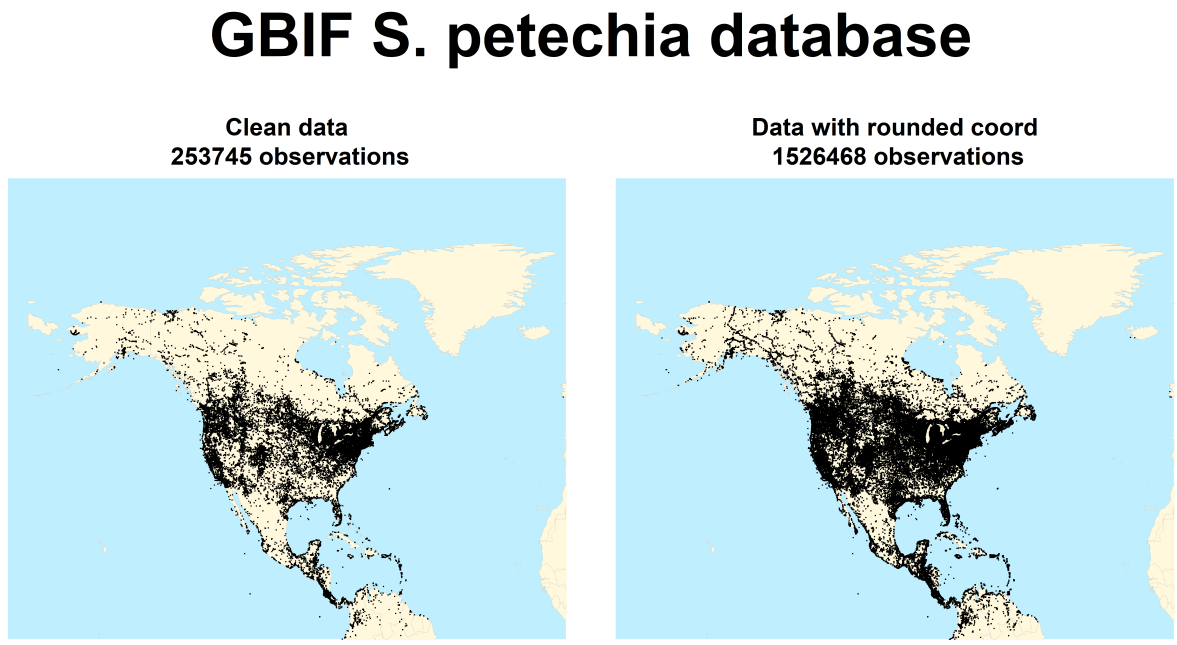


Figure 1: Data without issues vs the same data including also observations with rounded coordinates

-15°E, 8°N, 90°N.

```
# plot map with and without rounded coordinates
png("S_petechia_GBIF_maps.png", height=6, width=10,  units='in', res=600)
par(mfrow=c(1, 2),      # 1x2 layout
    oma=c(0, 0, 5, 0), # rows of text at the outer bottom left top right margin
    mar=c(3, 1, 3, 1), # space for row of text at ticks and to separate plots
    mgp=c(2, 1, 0))    # axis label at 2 rows distance, tick labels at 1 row

plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border="grey",
     xlim=c(-180,-15), ylim=c(8,90),
     main=paste("Clean data\n", dim(db1b)[1], "observations", sep=" "))#,
points(as.numeric(db1b$decimalLongitude), as.numeric(db1b$decimalLatitude), pch=".")

plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border="grey",
     xlim=c(-180,-15), ylim=c(8,90),
     main=paste("Data with rounded coord\n", dim(db2b)[1], "observations", sep=" "))#,
points(as.numeric(db2b$decimalLongitude), as.numeric(db2b$decimalLatitude), pch=".")

title(main="GBIF S. petechia database",cex.main= 3,
      outer=TRUE, line=2)

dev.off()
```

For the analyses only the summer (native breeding) range has been kept. From the Handbook of the Birds of the World website: "Eastern populations leave breeding grounds early, from mid-July, and move South on broad front through North America and return migration also early, reaching breeding grounds from early April in South, late May in far North. Western populations migrate a few weeks later, in both autumn and

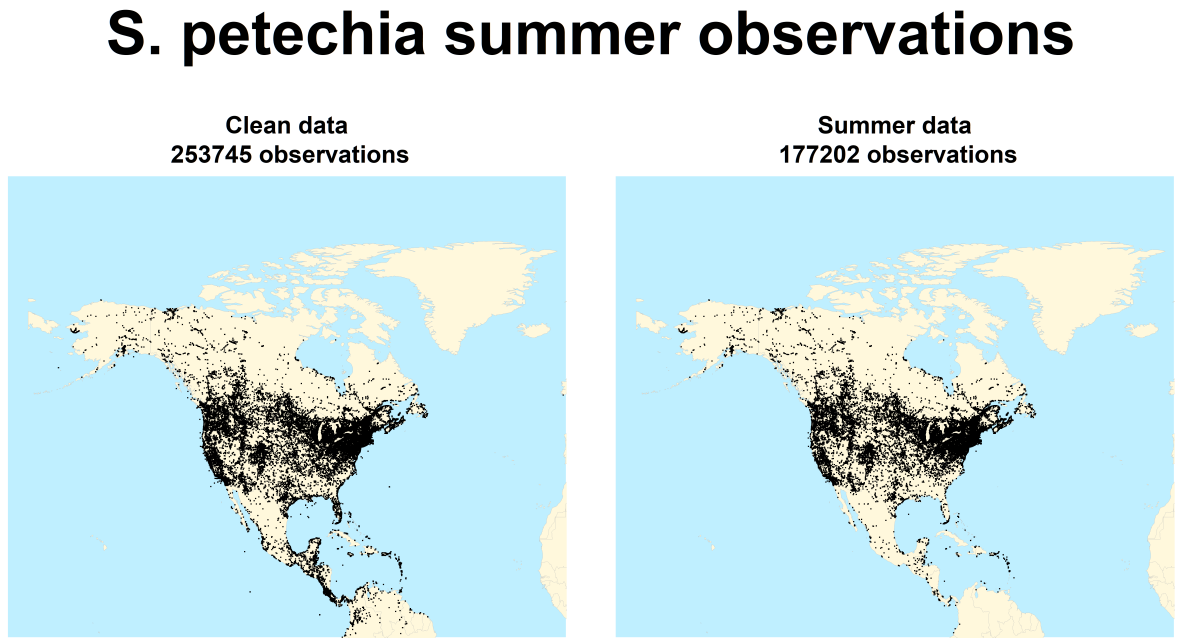


Figure 2: All data VS data only collected during summer

spring."

A first filter has been applied based on the month of the observation, considering the narrower temporal range described above

```
months <- c(5,6,7)
sum2 <- db1b[db1b$month %in% months,]

png("S_petechia_allVSsummer_maps_2.png", height=6, width=10,  units='in', res=600)
par(mfrow=c(1, 2),      # 1x2 layout
    oma=c(0, 0, 5, 0), # rows of text at the outer bottom left top right margin
    mar=c(3, 1, 3, 1), # space for row of text at ticks and to separate plots
    mgp=c(2, 1, 0))    # axis label at 2 rows distance, tick labels at 1 row

plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border="grey",
     xlim=c(-180,-15), ylim=c(8,90),
     main=paste("Clean data\n", dim(db1b)[1], "observations", sep=" "))
points(as.numeric(db1b$decimalLongitude), as.numeric(db1b$decimalLatitude), pch=".")

plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border="grey",
     xlim=c(-180,-15), ylim=c(8,90),
     main=paste("Summer data\n", dim(sum2)[1], "observations", sep=" "))
points(as.numeric(sum2$decimalLongitude), as.numeric(sum2$decimalLatitude), pch=".")

title(main="S. petechia summer observations",cex.main= 3,
      outer=TRUE, line=2)

dev.off()
```

Such filtering reduced the database to 177202 observations. The plot still shows observations outside the expected native breeding range for the species, as reported in the website BirdLife.

It is then necessary to clean the dataset based the provided masks. For this task it is better to include the widest temporal range for the native breeding (from April to May), and remove duplicates (observations with the same latitude and longitude), in order to reduce computing time afterwards, as the models used do not consider frequencies.

```
months <- c(4,5,6,7)
sum1 <- db1b[db1b$month %in% months,]

pts <- sum1[,c("gbifID","decimalLatitude","decimalLongitude")]
pts <- pts[!duplicated(pts[,c("decimalLatitude","decimalLongitude")]),]
```

The resulting database (50587 observations) has been then remapped based on the grid of the climate files used later in the analysis, and again any duplicate has been removed.

```
library(ncdf4)

# Function to modify coordinates
wherenearest <- function(val,matrix) {
  dist=abs(matrix-val)
  index=which.min(dist)
  return( index )
}

# Environmental variables for the whole area, in the present
envdata <- read.table("NAmerica/EnvirVar/EnvirVar_NAmerica_0.txt",header=TRUE, sep=" ")
colnames(envdata)[1:2] <- c("long","lat")

#Extract latitude and longitude
lon <- as.vector(unique(envdata$long))
lat <- as.vector(unique(envdata$lat))

# modify latitude and longitude to match the grid of the climatic reconstructions
pts$decimalLatitude <- sapply(pts$decimalLatitude, function(x)
  lat[wherenearest(x,lat)])
pts$decimalLongitude <- sapply(as.numeric(pts$decimalLongitude), function(x)
  lon[wherenearest(x,lon)])

# remove duplicates
pts <- pts[!duplicated(pts[,c("decimalLatitude","decimalLongitude")]),]
```

The dataset, reduced to 4315 observations, is now ready to cleaned based on the native breeding range mask. This task takes significantly more time whit much larger datasets, this is why some of the filtering has been done beore this step.

```
library(PBSmapping)
library(rgeos)
library(maptools)

# read mask data
myspecies<-readShapePoly("Setophaga_petechia",
                         proj4string=CRS("+proj=longlat +datum=WGS84"))

sapply(slot(myspecies, "polygons"), function(x) slot(x, "ID"))
```

```
# subset it for presence=1 or 2
myspecies.sub<-myspecies[myspecies$PRESENCE<3,]
myspecies.sub<-myspecies[myspecies$ORIGIN<3,]

# select the summer component
myspecies<-myspecies.sub[myspecies.sub$SEASONAL=="2",]

# create Polysets for summer and resident component:
# merging all polygons to create a single PID, needed for later operations
myspecies<-SpatialPolygons2PolySet(myspecies)
if (length(unique(myspecies$PID))>1) {
  myspecies2<-joinPolys(myspecies,operation="UNION")
}
myspecies<-PolySet2SpatialPolygons(myspecies)

# Subset data if within polygon
# define coordinates
xy <- pts[,c(3,2)]

#transform into SpatialPointsDataFrame
df <- SpatialPointsDataFrame(coords=xy, data=pts[,1],
                             proj4string=myspecies@proj4string)

# keep only points in shapefile
pts <- df[!is.na(over(df,myspecies)),]
```

After this last filtering (leaving 3364 observations) it is possible to create the input file for the following steps. The file must have three columns, two for the geographical coordinates and a third one with a 1 for the presences (and, if needed, 0 for the absences).

```
distrib <- cbind(pts$decimalLongitude,
                 pts$decimalLatitude,
                 rep(1, length(pts$decimalLongitude)))
colnames(distrib) <- c("long","lat","S.petechia")


# save table
write.table(distrib, file="S_petechia_distrib.txt", quote=FALSE, sep="\t", row.names=FALSE)

# plot
png("S_petechia_final_dataset.png", height=6, width=10,  units='in', res=600)
par(mfrow=c(1, 2),      # 1x2 layout
    oma=c(0, 0, 5, 0), # rows of text at the outer bottom left top right margin
    mar=c(3, 1, 3, 1), # space for row of text at ticks and to separate plots
    mgp=c(2, 1, 0))    # axis label at 2 rows distance, tick labels at 1 row

plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border="grey",
     xlim=c(-180,-15), ylim=c(8,90),
     main=paste("Final data\n", dim(pts)[1], "observations", sep=" "))
points(as.numeric(pts$decimalLongitude),
       as.numeric(pts$decimalLatitude),
       pch=".", col="red")

plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border="grey",
```

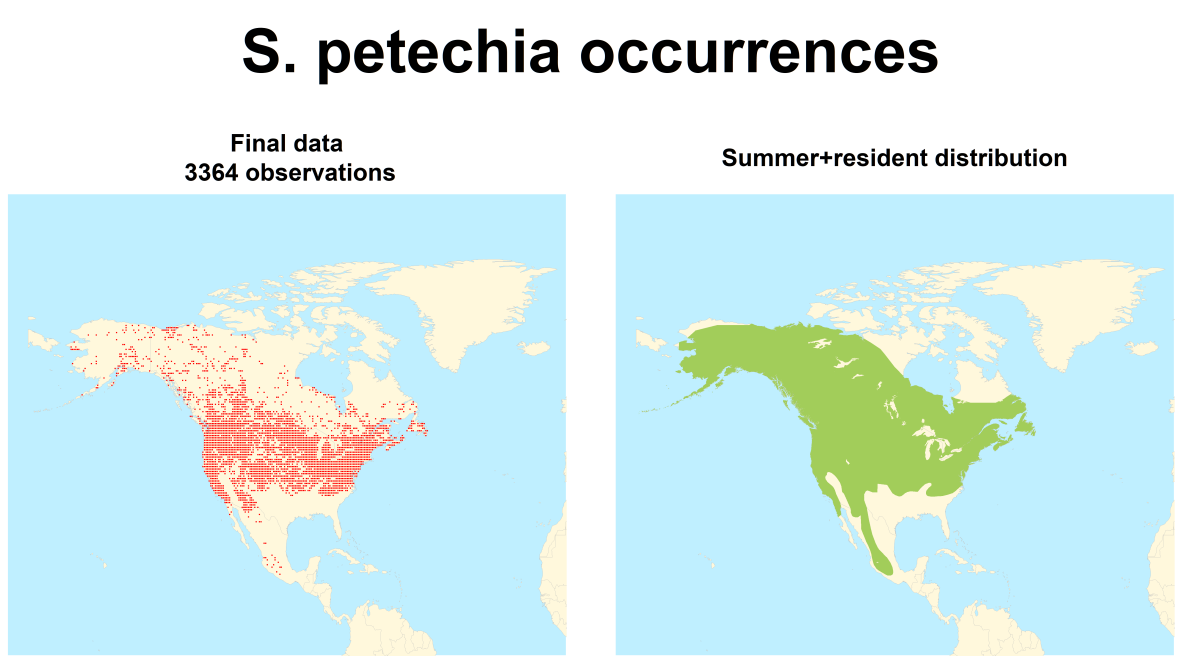


Figure 3: Final dataset

```
     xlim=c(-180,-15), ylim=c(8,90), main="Summer+resident distribution")
plot(myspecies, add=TRUE, col="darkolivegreen3", border=NA)

title(main="S. petechia occurrences",cex.main= 3,
      outer=TRUE, line=2)

dev.off()
```

Ecological analyses

Open the species data (if not loaded already).

```
# Species data
distrib <- read.table("S_petechia_distrib.txt", header=TRUE, sep="\t")
```

Remove NAs from the already loaded table listing the environmental variables (no-land cells) and extract the environmental variables for the observation locations.

```
# Remove NA
envdata <-  envdata[complete.cases(envdata), ]

# Extract vars for observations
obs <- merge(distrib[,c(1,2)], envdata)

#head(obs)
```

Please note that the final number of observations is reduced at 3281 after removing the ones falling in no-land cells due to regridding.

Principal component analysis

Principal Component Analysis (PCA) based on the environmental variables.

```
# merge data in a table, last column distinguish between baseline and observations
envdata[,"set"] <- "baseline"
obs[,"set"] <- "obs"
db <- rbind(envdata,obs)

# select only climatic variables
dbMDS <- db[,3:12]

# vector of colors
col <- c(rep("black", dim(envdata)[1]),rep("yellow3",dim(obs)[1]))

prin_comp <- prcomp(dbMDS, center=TRUE,
                    scale.=TRUE)
#extract coordinates
ind.coord <- prin_comp$x

# Eigenvalues
eig <- (prin_comp$sdev)^2
# Variances in percentage
variance <- eig*100/sum(eig)
# Cumulative variances
cumvar <- cumsum(variance)
eig2 <- data.frame(eig=eig, variance=variance,cumvariance=cumvar)

#plot PC1 PC2
png("S_petechia_PCA.png", height= 7, width=8.5,  units='in', res=600)
plot(ind.coord[,1],ind.coord[,2], col=col, pch=20,
     ylab=paste("PC2 (",round(eig2[2,"variance"],2)," %)",sep=""),
     xlab=paste("PC1 (",round(eig2[1,"variance"],2)," %)",sep=""),
     main="PCA environmental variables")
legend("bottomright", c("Baseline","S.petechia"),pch=20,
       col=c("black","yellow3"), bty="n")
dev.off()
```

Plot the direction of each variable in the PCA space.

```
#plot direction variables
var_cor_func <- function(var.loadings, comp.sdev){
  var.loadings*comp.sdev
}

# Variable correlation/coordinates
loadings <- prin_comp$rotation
sdev <- prin_comp$sdev
var.coord <- var.cor <- t(apply(loadings, 1, var_cor_func, sdev))

#head(var.coord[, 1:4])

a <- seq(0, 2*pi, length=100)

png("S_petechia_direzPCA.png", height=7,width=7, units='in', res=600)
plot( cos(a), sin(a), type='l', col="gray",
```

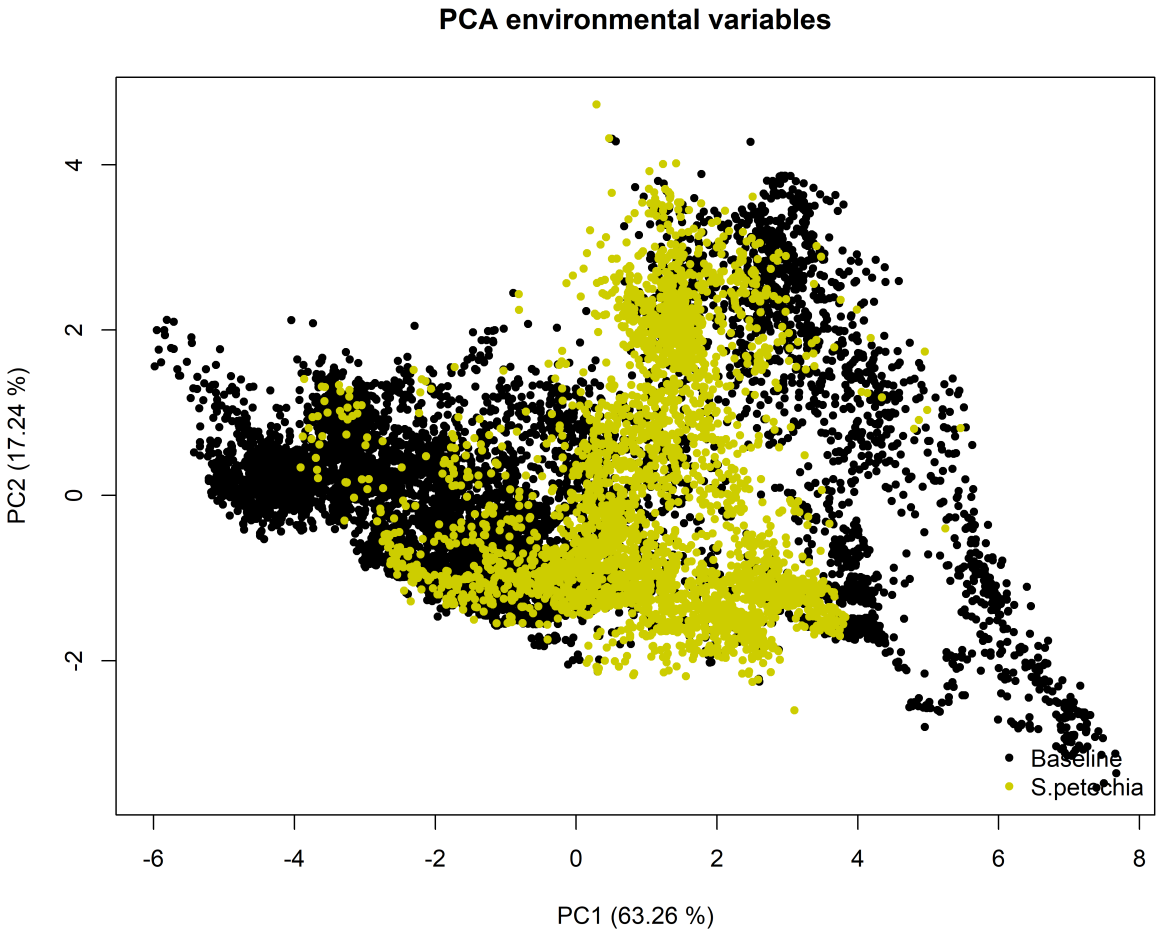


Figure 4: Principal Component Analysis (PCA) based on the environmental variables.

```
      xlab="PC1",  ylab="PC2")
abline(h=0, v=0, lty=2)
# Add active variables
arrows(0, 0, var.coord[, 1], var.coord[, 2],
       length=0.1, angle=15, code=2)
# Add labels tmin tmax totprec npp
text(var.coord, labels=rownames(var.coord), cex=1, adj=1, pos=3)
dev.off()
```

Variable distribution

Create a multiplot to compare the distribution of each variable in North America (black, on the left) with the distribution in the cells where the Yellow Warbler has been observed (orange, on the right).

```
library(beanplot)

# plot
png("S_petechia_variables.png", height=9, width=8,  units='in', res=600)
par(mfrow=c(4, 5),
    oma=c(0, 1, 5, 0),
    mar=c(1, 1, 3, 1),
    mgp=c(0.5, 0.5, 0))

# for each environmental variable
for (x in c(3:length(vars)+2)){

  # Beanplot distributions
  beanplot(db[db$set=="baseline",x],db[db$set=="obs",x],
         bw="nrd",side="both", col=list("black","yellow3"),
         border=c("black", "yellow3"), what=c(1,1,1,0),
         main=vars[x-2],xaxt='n')
}
title(main="Setophaga petechia climatic variables",cex.main= 3,
      outer=TRUE, line=2)
dev.off()
```

Cross-correlation

Based on the above plot and the PCA the most promising climate variables appear to be BIO1, BIO6, BIO7, BIO8, BIO9, BIO13, BIO14, BIO18, NPP and LAI. It is now necessary to calculate the cross correlation between them in order to exclude highly correlated ones.

The panel.cor function has been provided by Raquel A. Garcia, Stellenbosch University, South Africa.

```
# function to plot (by Raquel A. Garcia, Stellenbosch University, South Africa)
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr=c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

  test <- cor.test(x,y)
```

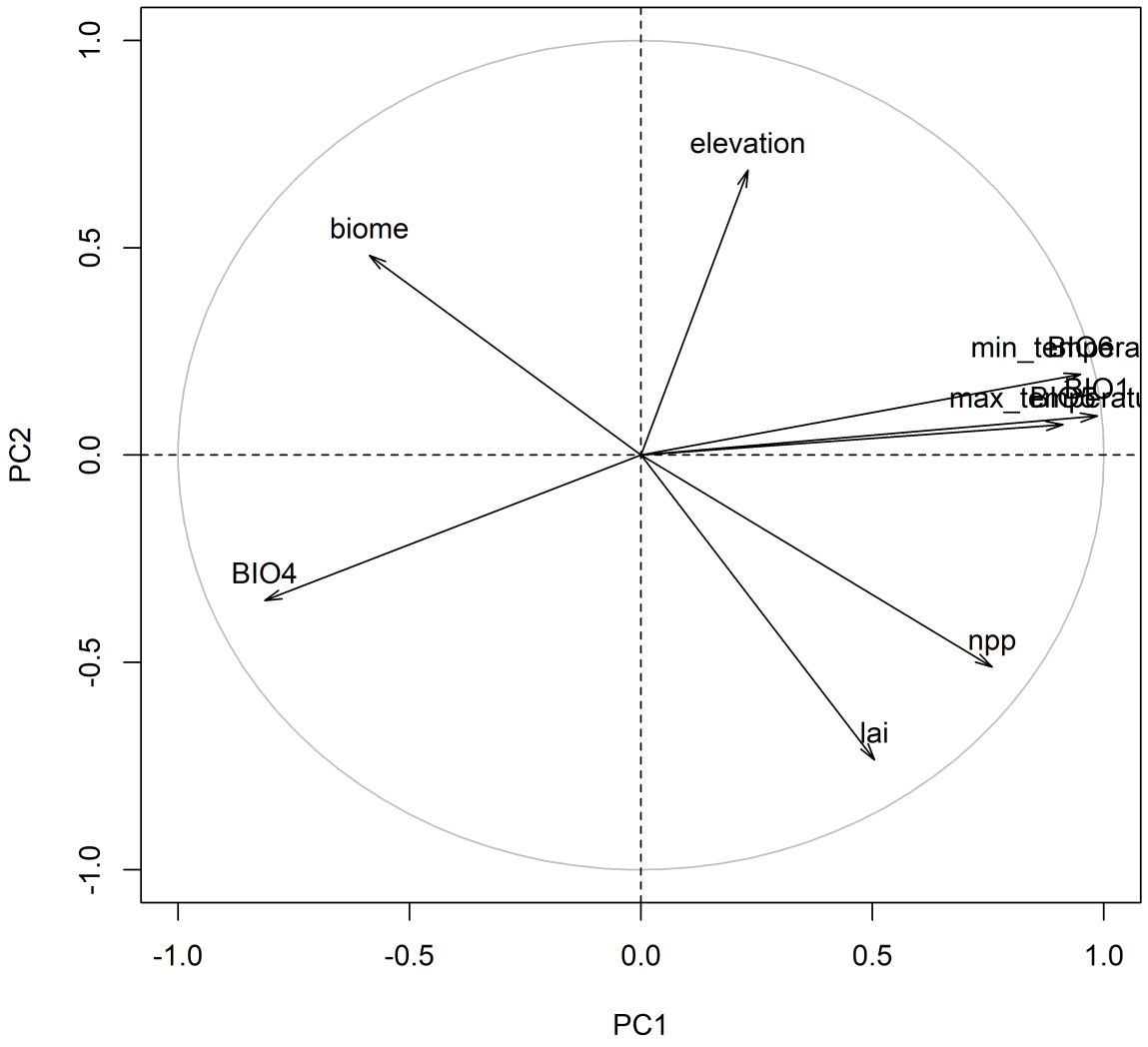


Figure 5: Plot of the direction of each variable in the PCA space.

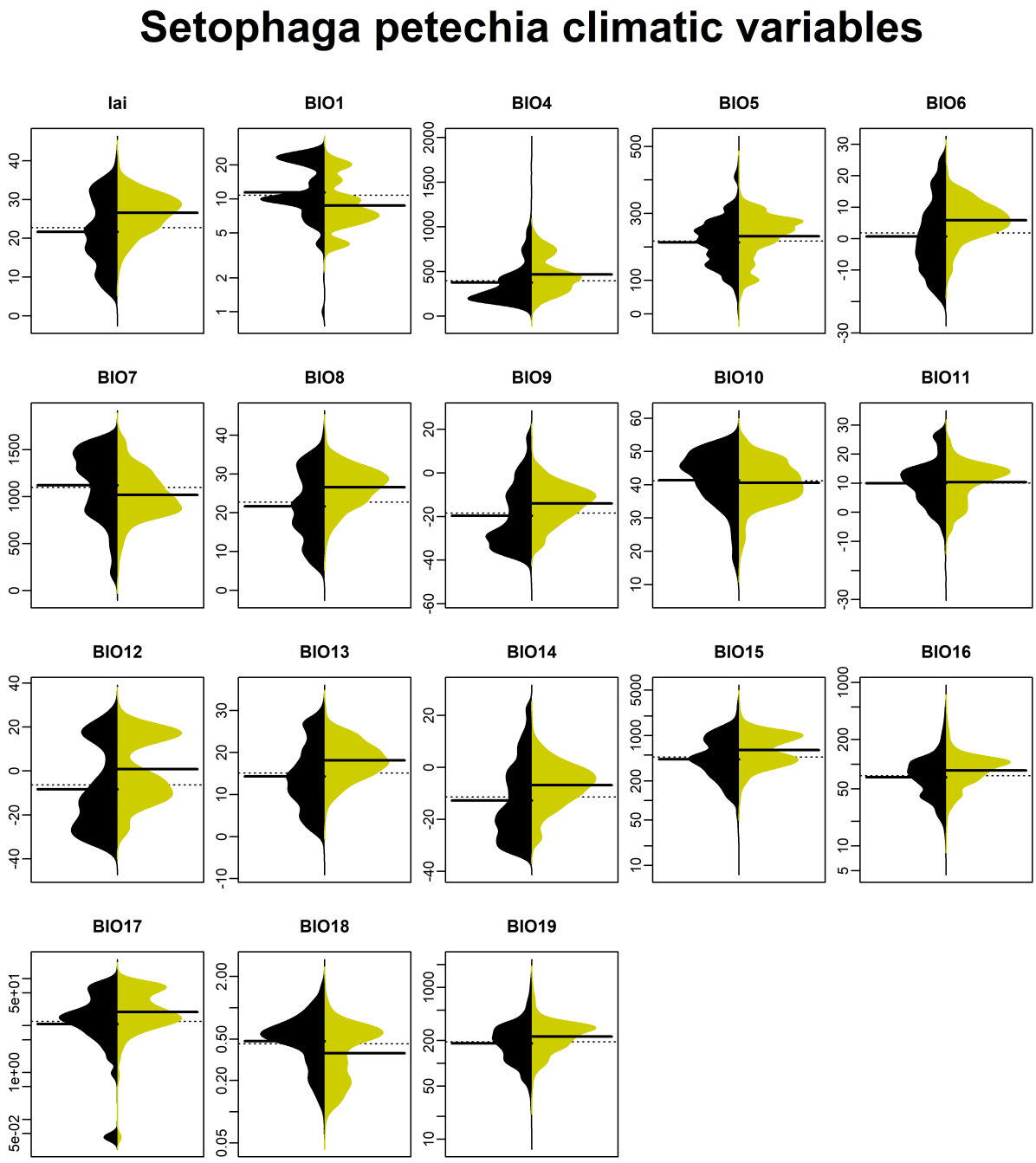


Figure 6: Comparison of the distribution of each variable in North America (black, on the left) with the distribution in the cells where the Yellow Warbler has been observed (yellow, on the right).

```
  # borrowed from printCoefmat
  Signif <- symnum(test$p.value, corr=FALSE, na=FALSE,
                   cutpoints=c(0, 0.001, 0.01, 0.05, 0.1, 1),
                   symbols=c("***", "**", "*", ".", " "))

  text(0.5, 0.5, txt, cex=cex * r)
  text(.8, .8, Signif, cex=cex, col=2)
}

# Variables of interest
ch <-c("npp","lai","BIO1","BIO6","BIO7","BIO8","BIO9",
         "BIO13","BIO14","BIO18")
# Pairwise correlation matrix
cormat <- cor(envdata[,ch])


# Plot correlation for variables of interest
png(filename="S_petechia_correlation_vars_interest.png",
    width=1200, height=900)
pairs(envdata[,ch], lower.panel=panel.smooth, upper.panel=panel.cor)
dev.off()
```

In order to reduce cross-correlation we only considered variables correlated up to 0.7.

```
library(caret)

# define highly correlated variables
hicor <- findCorrelation(cormat, cutoff=0.7)

# plot correlation for variables with correlation below 0.7
png(filename="S_petechia_correlation_uncorr_vars.png",
    width=1200, height=900)
pairs(envdata[,ch[-c(hicor)]], lower.panel=panel.smooth, upper.panel=panel.cor)
dev.off()
```

Species distribution modelling

The following chapter details the Species Distribution modelling on the basis of the observed presences of the species, and climatic reconstructions.

Preparing the environment: load libraries, set directories for the input files and define other parameters.

```
library(RColorBrewer)

# Raster maps for climatic variables through time (NAmerica)
basefile <- "NAmerica/"
# Directory where oputput is stored
BiomodData <- "BiomodData/"

# Species name (can be also loaded from the distribution file - 3rd column name)
sp <- "S.petechia"
# uncorrelated variables of interest
vars1 <- ch[-c(hicor)]
# Color palette
c1 <- colorRampPalette(c("khaki1","yellow3"))(6)
```

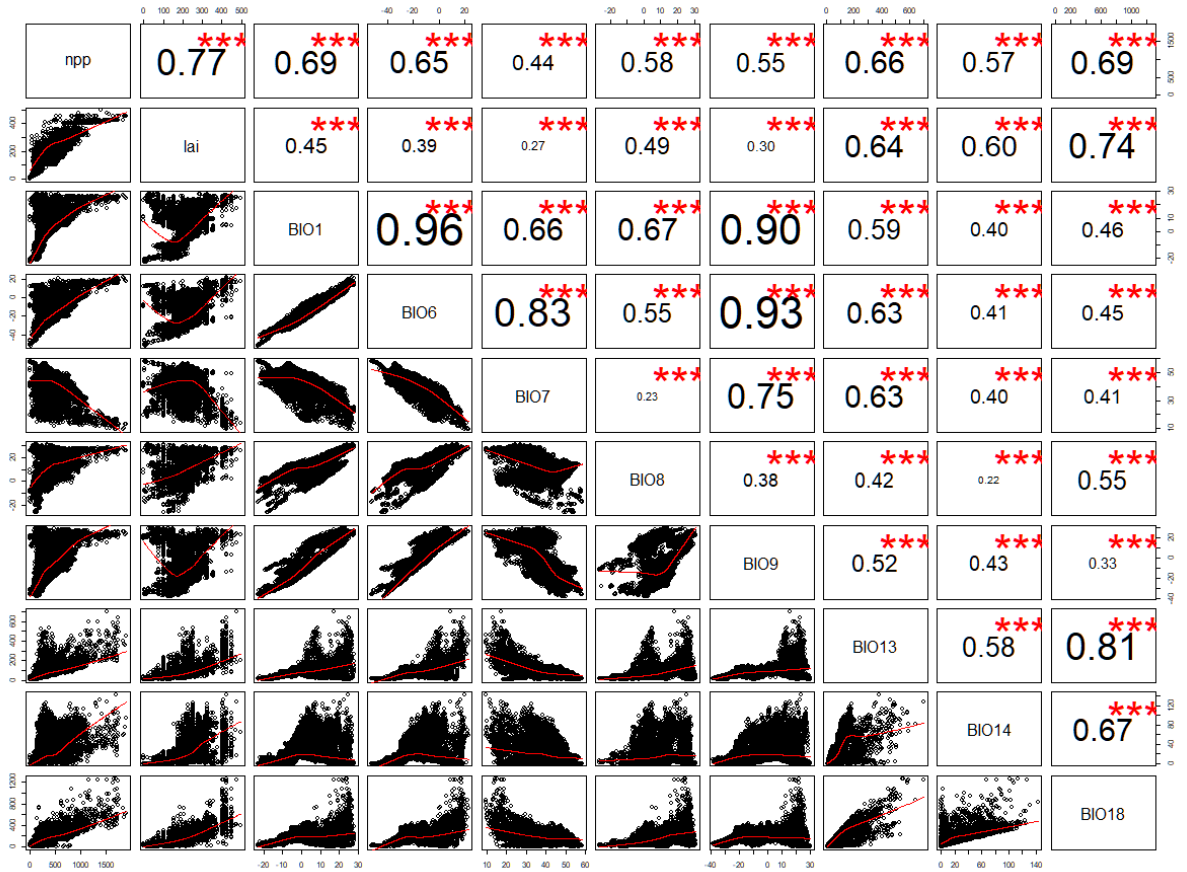


Figure 7: Correlation between all climatic variables of interest.

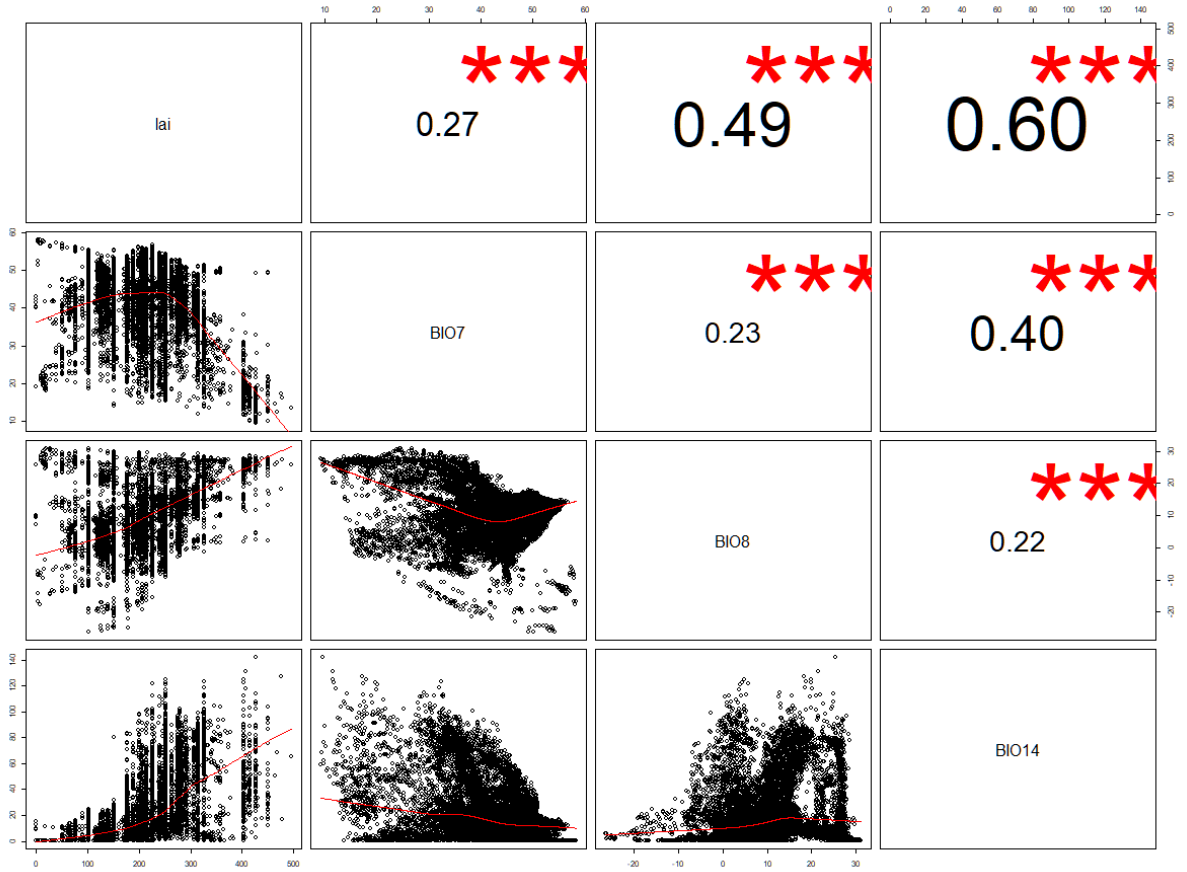


Figure 8: Correlation between chosen uncorrelated climatic variables (threshold=0.7).

```
c2 <- colorRampPalette(c("yellow3","black"))(8)
cols <- c(c1[2:5],c2[2:8])
```

Please be aware that if the species name includes an underscore (e.g.*S_petechia*) it will be transformed into a point ("*S.petechia*") by the program itself when using it within the output file names.

Spatial thinning

In order to reduce the geographic bias associated to an uneven geographic sampling of the species we decided to thin the dataset based on a minimum distance of 70 km based on 100 repetitions with the R package SpThin.

```
library(spThin)

# Create output directory
dir.create(file.path("Thin"))

# Spatial thinning
t <- thin(loc.data=as.data.frame(distrib),
          lat.col="lat", long.col="long", spec.col="S.petechia",
          thin.par=70, reps=100, locs.thinned.list.return=TRUE,
          write.files=TRUE, max.files=10, out.dir="Thin/",
          out.base="S.petechia", write.log.file=FALSE)

t_dist <- read.table("Thin/S.petechia_thin1.csv",header=TRUE,  sep=",")
#t_dist <- read.table("Thin_100km/S.petechia_thin1.csv",header=TRUE,  sep=",")
t_dist <- t_dist[,c(2,3,1)]

# plot thinning
png( "S_petechia_thinned_dataset.png", height=6, width=13,  units='in', res=600)
par(mfrow=c(1, 2),     # 1x2 layout
    oma=c(0, 0, 5, 0), # rows of text at the outer bottom left top right margin
    mar=c(3, 1, 3, 1), # space for row of text at ticks and to separate plots
    mgp=c(2, 1, 0))    # axis label at 2 rows distance, tick labels at 1 row

plot(newmap, col="cornsilk", bg="lightblue1", border="grey", xlim=c(-180,-15), ylim=c(8,90),
     lwd=0.05, main=paste("Dataset\nN =", dim(distrib)[1], sep=" "))
points(as.numeric(distrib[,"long"]),as.numeric(distrib[,"lat"]), pch="*", col="yellow4")

plot(newmap, col="cornsilk", bg="lightblue1", border="grey", xlim=c(-180,-15), ylim=c(8,90),
     lwd=0.05, main=paste("Thinned dataset\nN =", dim(t_dist)[1], sep=" "))
points(as.numeric(t_dist$long), as.numeric(t_dist$lat), pch="*", col="black")

title(main="S. petechia occurrences",cex.main= 3,
      outer=TRUE, line=2)

dev.off()
```

Geographic input file

Formatting the climate data as required for the modelling step. Loading modern day climate data (as they start with "*0_*", meaning present-day) and extracting rasters for selected variables for the formatting function.

```
library(raster)
library(rgdal)
```

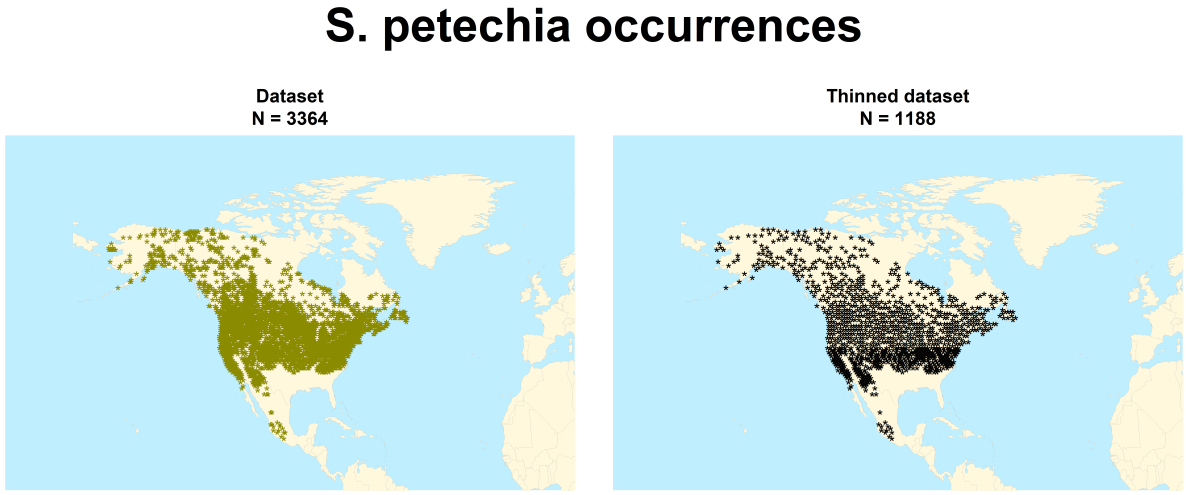


Figure 9: Original (left) and spatially thinned (right) dataset. The threshold used for spatial thinning is 70 km.

```
library(SDMTools)

# environmental variables for modelling
expl.var <- stack()

# create raster stack
for(v in 1:length(vars1)){
  r <- raster(paste(basefile, vars1[v], "/NAmerica_",vars1[v],"_0.grd", sep=""), RAT=FALSE)
  expl.var <- stack( expl.var, r)
}
```

Pseudo-absence selection

We decided to randomly draw pseudo-absences from all possible points outside the original mask, with a sample size that equals the presences. As a first step the following code identifies the points available for pseudoabsences

```
# background data
bg <- data.frame(cbind(envdata[,1:2], S.petechia=rep(0,dim(envdata)[1])))

# define coordinates
xy <- bg[,1:2]

#transform into SpatialPointsDataFrame
df <- SpatialPointsDataFrame(coords=xy, data=bg,
                             proj4string=myspecies@proj4string)

# remove points within the mask
bg.clean <- df[is.na(over(df,myspecies)),]
bg.clean <- data.frame(bg.clean[,1:3])

# # plot
```

```
# plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border = "grey",
#      xlim = c(min(lon), max(lon)), ylim = c(min(lat), max(lat)))
# points(as.numeric(bg.clean[,"long"]), as.numeric(bg.clean[,"lat"]), pch="*", col="red")
```

The following code randomly drawn pseudoabsences 5 times and creates a table with information on pseudoabsences in the format of "PA.table" to be read by biomod2.

```
# number of absences and presences
pres <- dim(t_dist)[1]
abs <- dim(t_dist)[1]

# Absences table
abs.table <- data.frame(matrix("FALSE",abs*5,7), stringsAsFactors=FALSE)
colnames(abs.table) <- c("long","lat",paste("RUN",c(1:5),sep=""))

# variable to define the first line to be written
start <- 1

for (i in 1:5){
  ix <- sample(1:dim(bg.clean)[1], abs)
  abs.table[seq(start,abs*i),1:2] <- bg.clean[ix,1:2]
  abs.table[seq(start,abs*i),2+i] <- rep("TRUE",abs)
  start <- (abs*i)+1
}
```

Species input file

Formatting the species distribution and extracting the relevant information as required for the modelling step. This script is based on a species input file format with three columns, two for the geographical coordinates and a third one with a 1 for the presences.

```
#
resp.var <- as.numeric(c(rep(1, pres),rep(0,abs*5)))      # species presences + background
resp.xy <- rbind(t_dist[,c("long","lat")], abs.table[,c("long","lat")])# coordinates

# add presences to PA.table
pres.table <- data.frame(cbind(t_dist[,c("long","lat")]), matrix("TRUE",pres,5))
colnames(pres.table) <- c("long","lat",paste("RUN",c(1:5),sep=""))

# merge tables for presences and pseudoabsences
PA.table <- rbind(pres.table[,-c(1,2)], abs.table[,-c(1,2)])
PA.table[] <- lapply(PA.table, as.logical)
```

Creation of the input required for biomod2 using the already defined pseudo-absences.

```
library(biomod2)

biomodData <- BIOMOD_FormatingData(resp.var=resp.var,          # species distribution
                                   expl.var=expl.var,          # environmental variables
                                   resp.xy=resp.xy,            # coordinates of species
                                   resp.name=sp,               # name of the species column
                                   PA.strategy="user.defined",# user defined pseudo absences
                                   PA.table=PA.table,
                                   na.rm=TRUE)                 # do not consider NA
```

```
png("S_petechia_pseudoabsences_datasets.png", width=7, height=7, units='in', res=600)
plot(biomodData)
dev.off()
```

The resulting object can be stored on the computer executing the following code:

```
# create BiomodData directory
dir.create("BiomodData")
# store file
save(biomodData, file=paste("./BiomodData/",sp, sep=""))
```

Block cross-validation

The following code uses the package blockCV to split the dataset into two parts, one for calibration and one for evaluation. It is important to reload the dataset fro the biomodData object because formating the data may remove some points (because of NA in the variables associated).

We decided to split the data into 15 vertical (North-South) columns and creating 5 different datasets in which 12 of them are used for celibration and 3 for evaluation. The output of this process is saved as *DataSplitTable* the format to be read by biomod2.

```
library(blockCV)

# presences and absences taken from biomodData (with NA removed)
resp <- cbind(biomodData@coord,S.petechia=biomodData@data.species)
resp[] <- lapply(resp, as.numeric)

# transform into SpatialPointsDataFrame
resp <- SpatialPointsDataFrame(resp[,c("long", "lat")], resp, proj4string=crs(expl.var))

# create spatial blocks and saves plot
png("S_petechia_spatial_blocks.png", height=6, width=10,  units='in', res=600)
sb <- spatialBlock(speciesData = resp,
                   species = "S.petechia",
                   rasterLayer = expl.var,
                   rows = 15,
                   k = 5,
                   selection = "systematic",
                   #iteration = 500, # find evenly dispersed folds
                   biomod2Format = TRUE)
dev.off()

DataSplitTable <- sb$biomodTable
```

The following code allows to plot how the observations are used in the different runs.

```
# add coordinates to the table
dst.xy <- cbind(resp$long,resp$lat,DataSplitTable)
colnames(dst.xy)[1:2] <- c("x", "y")

# plot
png("S_petechia_blocks_CV.png", width=7, height=8, units='in', res=600)
par(mfrow=c(3, 2),
    oma=c(0, 0, 5, 0),
    mar=c(1, 1, 1, 1),
    mgp=c(2, 1, 0))
```

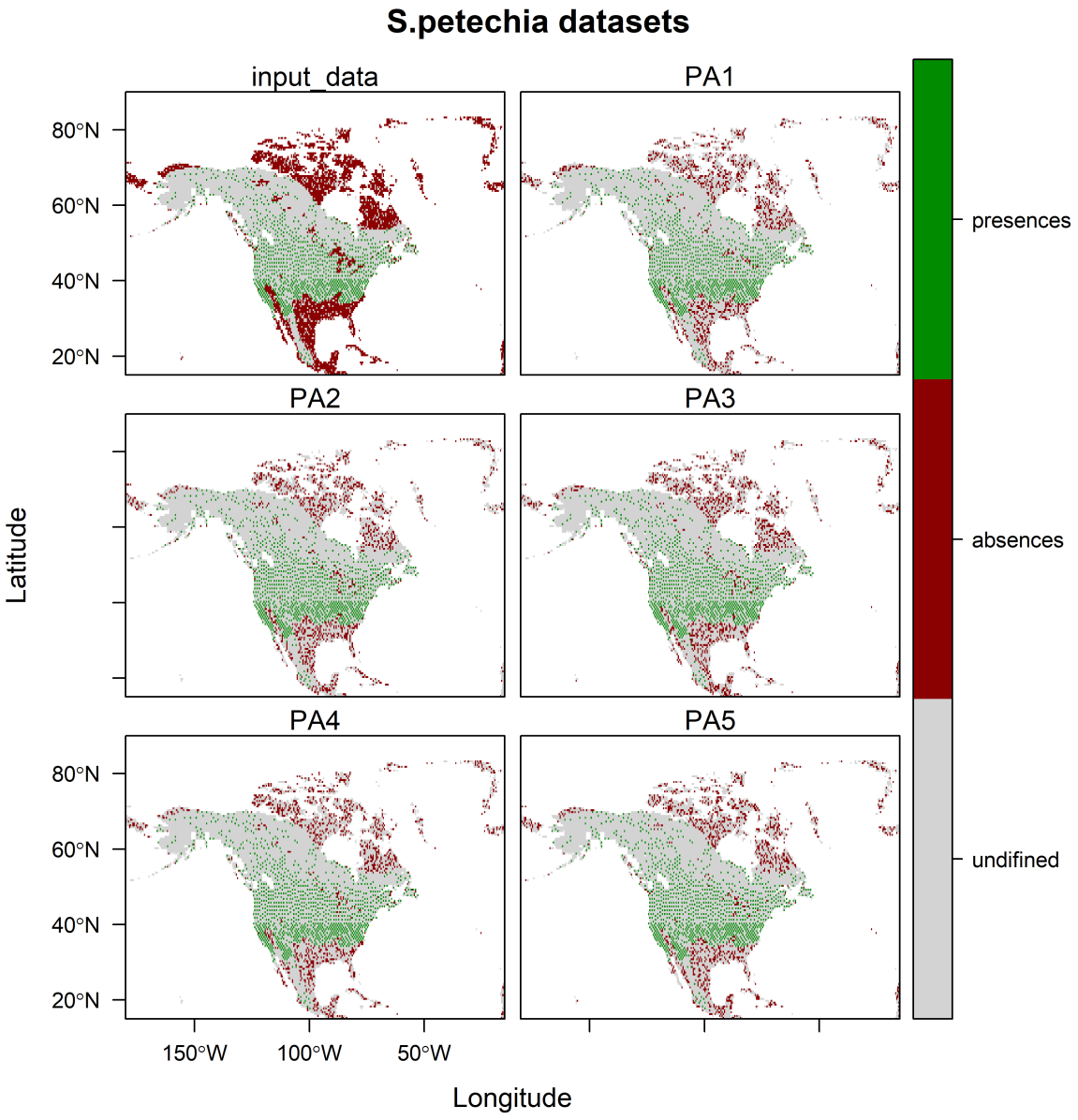


Figure 10: Pseudoabsences datasets (randomly drawn from the whole background in the same number as presences)

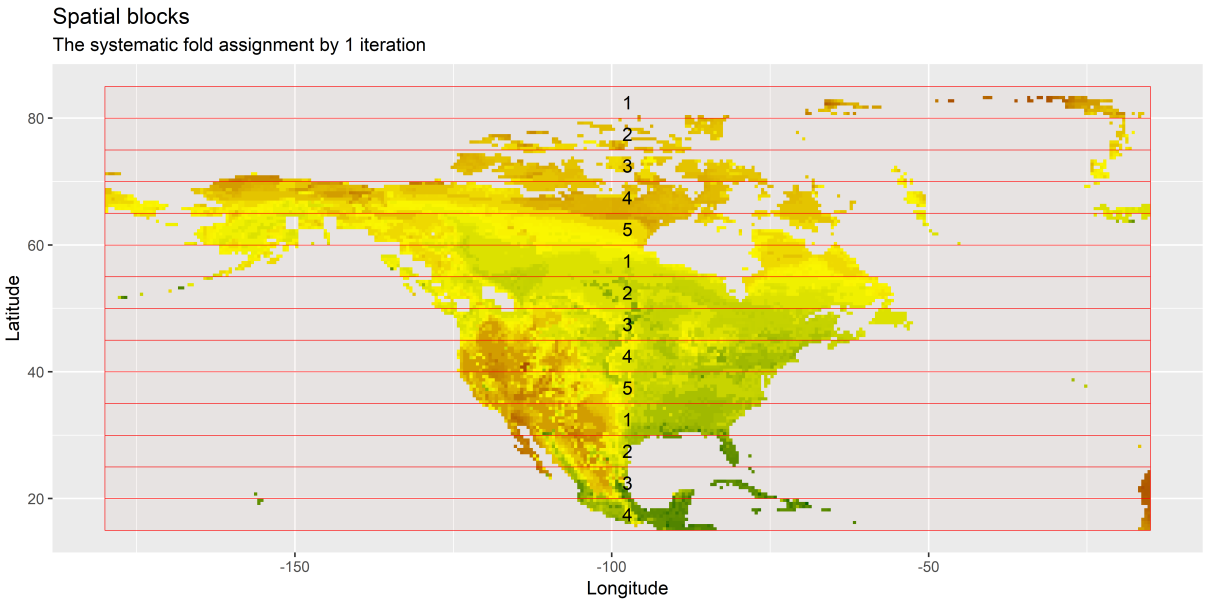


Figure 11: Spatial blocks defined by blockCV)

```
for (i in c(1:5)){

  col <- DataSplitTable[,paste("RUN",i,sep="")]
  col[col=="FALSE"] <- "yellow4"
  col[col=="TRUE"] <- "black"

  plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border = "grey",
       xlim = c(min(lon), max(lon)), ylim = c(min(lat), max(lat)), main=paste("RUN",i,sep=""))
  points(as.numeric(dst.xy[,"x"]), as.numeric(dst.xy[,"y"]), pch="*", col=col)
}

# title
title(main="S. petechia spatial blocks cross-validation", cex.main= 2, outer=TRUE, line=3)

dev.off()
```

Calibrating the models

The following code allows the calibration of the models using all algorithms available in biomod2, as specified by the vector. The data is split in two parts, 80% of the data are used for calibration and 20% for evaluation based on TSS. To define the model parameters it is possible to create an object with the default options with the command *BIOMOD_ModelingOptions()*.

```
ModOptions <- BIOMOD_ModelingOptions()

mods <- c("GLM","GBM","GAM","RF") # GBM = boosted regression tree
```

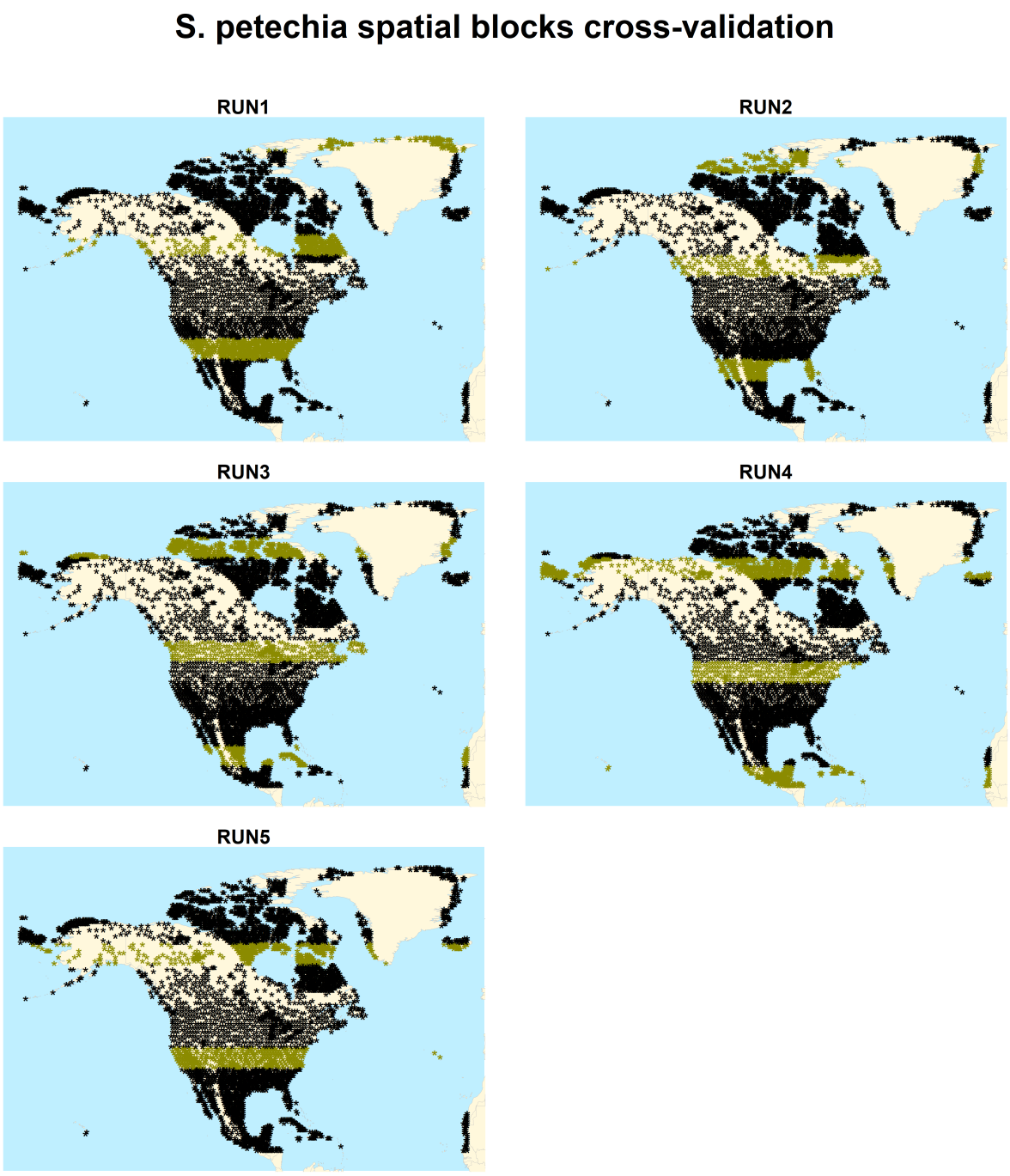


Figure 12: Spatial splitting of the data for each run: black for calibration, dark yellow for validation)

```
ModelOut <- BIOMOD_Modeling(biomodData,                    # input data
                            models=mods,                   # algorithms
                            models.options=ModOptions,     # options
                            DataSplitTable=DataSplitTable, # DataSplitTable issued by blockCV
                            #NbRunEval=3,                  # number of evaluations
                            #DataSplit=80,
                            models.eval.meth=c("TSS"),     # method for evaluat
                            SaveObj=T,
                            modeling.id=paste(sp,"_modelOut", sep=""),
                            do.full.models=FALSE,
                            rescal.all.models=T)
# model evaluations
modEval <- get_evaluations(ModelOut)

# TSS scores
write.table(modEval["TSS","Testing.data",,], "S.Petechia_models_eval.txt")
```

Projecting to the entire study area

Project the models to the entire study area (baseline period) using the same climate data used for the calibration.

```
# environmental information (same as used for calibration)
new.env <- expl.var

ProjBas <- BIOMOD_Projection(modeling.output=ModelOut, # model
                             selected.models="all",    # models to select
                             new.env= new.env,         # environmental data to project to
                             proj.name="Proj_NAmer",
                             binary.meth=c("TSS"),     # method for binary transformation
                             filtered.meth=c("TSS"),   # method to set values below zero
                             build.clamping.mask=T,    # out-of-calibration range
                             compress="xz")

# # make some plots sub-selected by str.grep argument
# png("S_petechia_proj_RF_RUN1.png", width=9, height=12, units='in', res=600)
# plot(ProjBas, str.grep = 'RF')
# dev.off()
```

Ensemble modelling

Defining the rules for generating ensembles and for evaluating them. The ensemble is built merging all algorithms together and is evaluated by TSS (threshold=0.7).

```
EnsMod <- BIOMOD_EnsembleModeling(modeling.output=ModelOut,
                                  chosen.models="all",      # models used
                                  em.by="all",              # by algorithm
                                  #VarImport=5,              # var importance permutations
                                  eval.metric=c("TSS"),     # evaluation metric
                                  eval.metric.quality.threshold=c(0.7),
                                  models.eval.meth=c("TSS"),# evaluation method
                                  prob.mean=T,
                                  prob.median=T,
                                  committee.averaging=T,
```

```
                                    prob.mean.weight=T,
                                    prob.mean.weight.decay="proportional")
```

Validation of the model

Get the evaluation scores for the ensemble model

```
eval <- get_evaluations(EnsMod)
write.table(eval, file="S_petechia_EnsembleEvaluation.txt")
```

Create a table with evaluation scores, sensitivity and specificity.

```
# change dimension names to make them more easily accessible
names(eval) <- c("mean_TSS","median_TSS","ca_TSS","wmean_TSS")

# Create a table with evaluation scores, sensitivity and specificity
out<- rbind(c(eval$mean_TSS[,1], eval$median_TSS[,1], eval$ca_TSS[,1], eval$wmean_TSS[,1]),
            c(eval$mean_TSS[,3], eval$median_TSS[,3], eval$ca_TSS[,3], eval$wmean_TSS[,3])/100,
            c(eval$mean_TSS[,4], eval$median_TSS[,4], eval$ca_TSS[,4], eval$wmean_TSS[,4])/100)
```

Plot in a single figure three barplots showing the evaluation scores based on TSS for each of the five independent evaluations.

```
# index for the maximum TSS
ix <- which.max(out[1,])

# Plot barplots of evaluation
colnames(out) <- c("Mean","Median","Committee\naverage","Weighted\nmean")
rownames(out) <- c("Evaluation score", "Sensitivity","Specificity")

png("S_petechia_evaluation_scores_ensemble.png", width=9, height=6, units='in', res=600)
layout(matrix(c(1,2), ncol=1, byrow=TRUE), heights=c(3.8,0.7))
par(oma=c(0, 0, 4, 0), # rows of text at the outer bottom left top right margin
    mar=c(2, 3, 3, 1))#, # space for row of text at ticks and to separate plots

barplot(t(out), beside=TRUE, col=cols[c(6,2,8,4)], border=NA, ylim=c(0,1))

par(mai=c(0,0,0,0))
plot.new()
legend("center", colnames(out), fill=cols[c(6,2,8,4)], border=NA, bty="n",ncol=4)

title(main=paste("Evaluation score (TSS) -", sp, sep=" "), cex.main=1.5, outer=TRUE, line=1)

dev.off()
```

Ensemble forecasting

The following code allows projecting the ensemble to build the consensus projections based on the same climatic variables already used for the other steps (new.env, which is a copy of expl.var)

```
EnsBas <- BIOMOD_EnsembleForecasting(EM.output=EnsMod,          # output from ensemble modelling
                                     projection.output=ProjBas, # output from projection
                                     selected.models="all",     # model chosen
                                     binary.meth=c("TSS"),      # method for binary transf
                                     filtered.meth=c("TSS"),    # method for filtering
                                     compress=T)                # compression
```

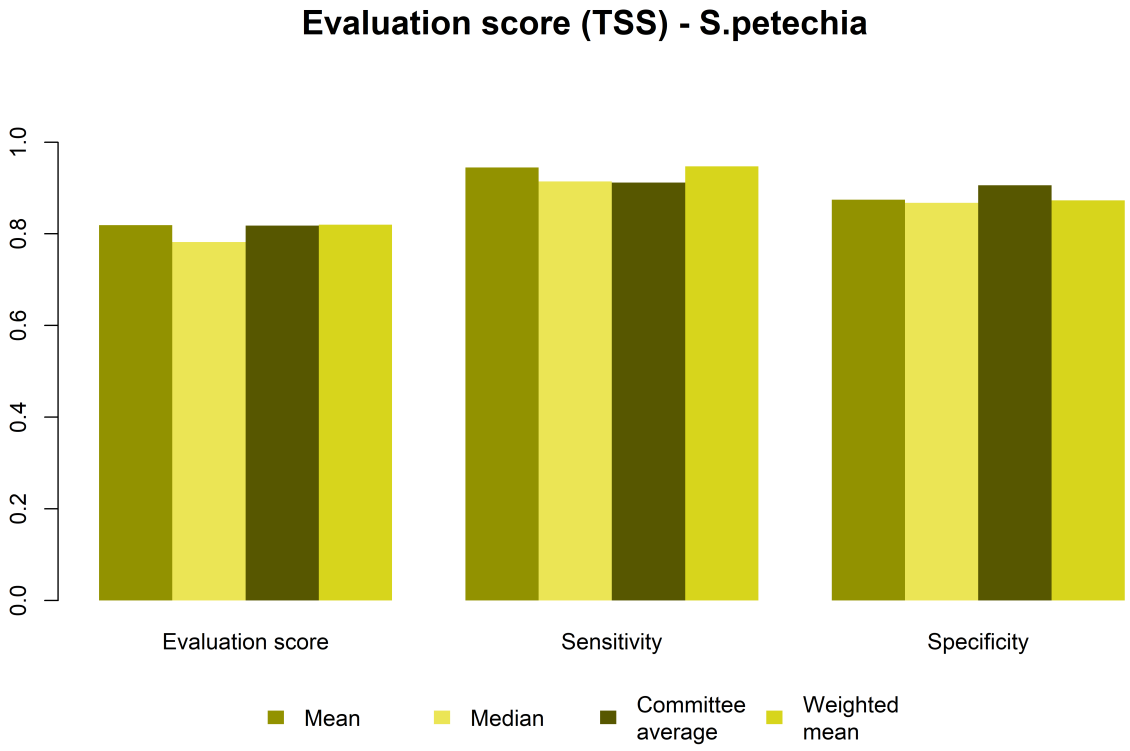


Figure 13: Ensemble model evaluation

Exploring the output: loading the binary consensus projections:

```
#list of files in directory
filelist <- list.files(paste(sp,"/proj_Proj_NAmer/", sep=""))
# selecting consensus projections
enslist <- filelist[grep("ensemble", filelist)]
enspbin <- stack(paste(sp,"/proj_Proj_NAmer/", enslist[2], sep=""))
```

and plotting them:

```
ens.methods <- c("Mean", "Median", "Committee average", "Weighted mean")

# Read ice mask
r <- raster('Ice/Mask_0.nc', var="ice_thickness")

png("S_petechia_ensemble_projection.png",
    width=12, height=9, units='in', res=600)
par(mfrow=c(2, 2),
    oma=c(0, 0, 0, 1),
    mar=c(3, 3, 2, 5),
    mgp=c(2, 1, 0))

for (x in 1:4) {
  #plot selected rasters (ensemble models)
  plot(raster(enspbin,c(x)), main=paste("Modern projection", ens.methods[x], sep=" - "), axes=FALSE,
       box=FALSE, colNA="lightblue1", col=c("cornsilk",terrain.colors(8)[7:1]))

  # plot observation points
  #points(resp.xy[,"long"], resp.xy[,"lat"], pch=".", cex=0.5)

  # plot ice
  plot(r, add=TRUE, col=c(NA,"honeydew3"), border=NA, useRaster=F, legend=FALSE)
}

dev.off()
```

Projection backwards in time

The following code repeat the whole process in order to project the potential distribution of the species backwards in time.

```
dir.create(file.path("Past"), showWarnings=FALSE)

# list of kyrs ago available
ref <- c(1:22, seq(24,50,2))

# for each kyrs
for (i in ref){
  # Read ice mask
  ice <- raster(paste("Ice/Mask_",i,".nc", sep=""), var="ice_thickness")

  # create raster stack
  past.env <- stack()

  # loading climate data for selected variables
  for(v in 1:length(vars1)){
```

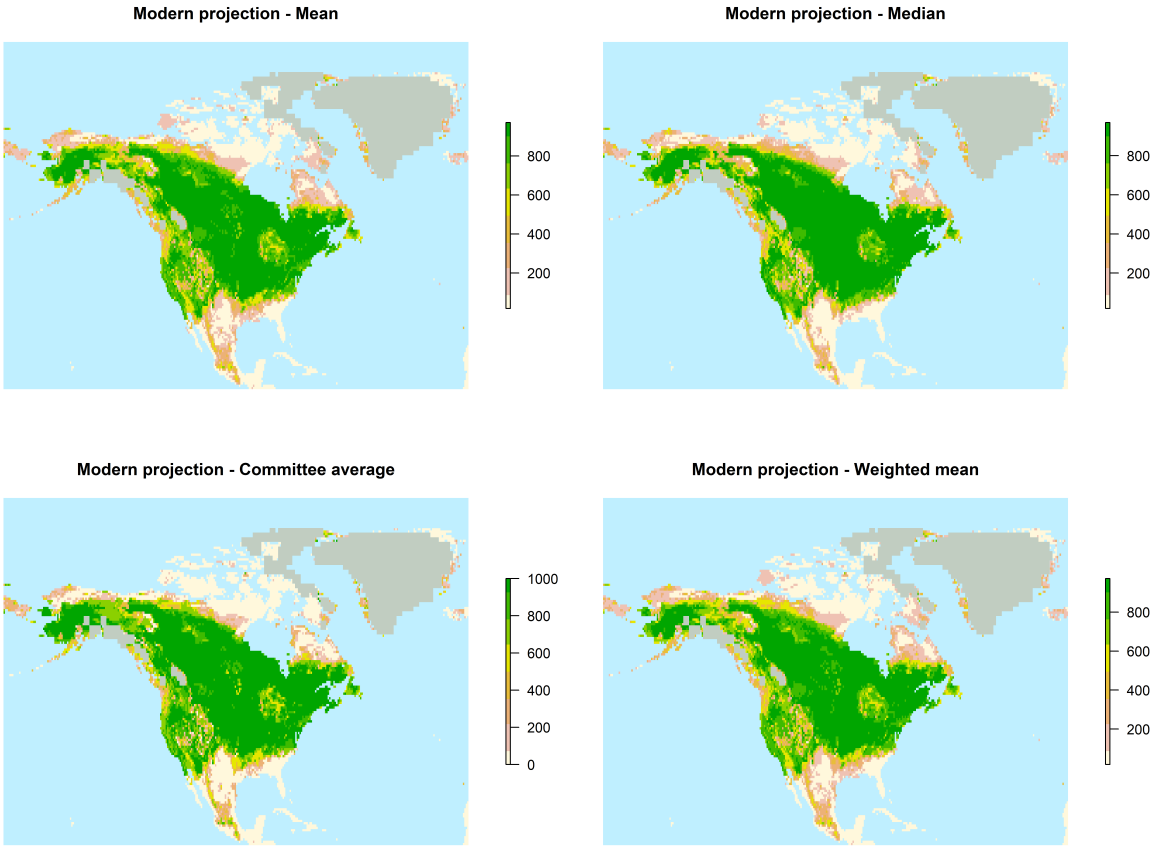


Figure 14: Plot of the ensemble projections

```
r <- raster(paste(basefile, vars1[v], "/NAmerica_",vars1[v],"_",i, ".grd", sep=""), RAT=FALSE)
past.env <- stack(past.env, r)
}

# Project model into the past
ProjPast <- BIOMOD_Projection(modeling.output=ModelOut,  # model
                              selected.models="all",         # model to select
                              new.env= past.env,          # environmental data to project to
                              proj.name=paste(i,"_NAmer", sep=""),
                              binary.meth="TSS",          # method for binary transformation
                              filtered.meth="TSS",        # method to set values below zero
                              build.clamping.mask=T,      # out-of-calibration range
                              compress="xz")

# Project ensemble to the past
EnsPast <- BIOMOD_EnsembleForecasting(EM.output=EnsMod,           # output from ensemble mod
                                      projection.output=ProjPast,  # output from projection
                                      selected.models="all",       # model chosen
                                      binary.meth=c("TSS"),        # eval method for bin transf
                                      filtered.meth=c("TSS"),      # eval method for filtering
                                      compress=T)                  # compression?


filelist <- list.files(paste(sp,"/proj_",i,"_NAmer/", sep="")) #list of files in directory
enslist <- filelist[grep("ensemble", filelist)]  # selecting consensus projections
enspbin <- stack(paste(sp,"/proj_",i,"_NAmer/", enslist[2], sep=""))

png(paste("Past/S_petechia",i,"all_ensemble_projections.png", sep="_"),
    width=12, height=9, units='in', res=600)

op<-par(mfrow=c(2, 2),
        oma=c(0, 0, 0, 1),
        mar=c(3, 3, 2, 5),
        mgp=c(2, 1, 0))

for (x in 1:4) {
  # plot selected rasters (ensemble models)
  plot(raster(enspbin,c(x)), main=paste("Projection",i,"k years ago -", ens.methods[ix], sep=" "),
       axes=FALSE, box=FALSE, colNA="lightblue1", col=c("cornsilk",terrain.colors(8)[7:1]))
  # add ice
  plot(ice, add=TRUE, col=c(NA,"honeydew3"), border=NA, useRaster=F, legend=FALSE)
  }
dev.off()


for (x in 1:4) {
  png(paste("Past/S_petechia",i,ens.methods[x],"ensemble_projection.png", sep="_"),
      width=12, height=9, units='in', res=600)
  plot(raster(enspbin,c(x)), main=paste(ens.methods[x], i, "k years ago", sep=" "))
  plot(ice, add=TRUE, col=c(NA,"honeydew3"), border=NA, useRaster=F, legend=FALSE)
  dev.off()
}
```

```
}
```

The following code creates a plot showing the mask, the data and the projection of the ensemble in four key periods: present-day, middle Holocene (6kyrs ago), Last Glacial Maximum (21 kyrs ago) and 50 kyrs ago. In order to create this plot we used the Mean of the ensemble, which (togheter with the weighted mean) showed the best fit.

```
# read raster modern
filelist <- list.files("S.petechia/proj_Proj_NAmer/") #list of files in directory
enslist <- filelist[grep("ensemble", filelist)]  # selecting consensus projections
enspbin <- stack(paste("S.petechia/proj_Proj_NAmer/", enslist[2], sep=""))
pr <- raster(enspbin,1)

# Holocene
filelist <- list.files("S.petechia/proj_6_NAmer/")
enslist <- filelist[grep("ensemble", filelist)]
enspbin <- stack(paste("S.petechia/proj_6_NAmer/", enslist[2], sep=""))
HOL <- raster(enspbin,1)

# LGM
filelist <- list.files("S.petechia/proj_21_NAmer/")
enslist <- filelist[grep("ensemble", filelist)]
enspbin <- stack(paste("S.petechia/proj_21_NAmer/", enslist[2], sep=""))
LGM <- raster(enspbin,1)

# 50k
filelist <- list.files("S.petechia/proj_50_NAmer/")
enslist <- filelist[grep("ensemble", filelist)]
enspbin <- stack(paste("S.petechia/proj_50_NAmer/", enslist[2], sep=""))
k50 <- raster(enspbin,1)

# Ice
r <- raster('Ice/Mask_0.nc', var="ice_thickness")
r12 <- raster('Ice/Mask_12.nc', var="ice_thickness")
r21 <- raster('Ice/Mask_21.nc', var="ice_thickness")
r50 <- raster('Ice/Mask_50.nc', var="ice_thickness")

# plot
png(paste("S_petechia_6_plot.png",sep="_"), height=4.5, width=9,  units = 'in', res = 600)
par(mfrow = c(2, 3),     # 2x2 layout
    oma = c(0, 0, 5, 2), # rows of text at the outer bottom left top right margin
    mar = c(1, 2, 1, 3), # space for row of text at ticks and to separate plots
    mgp = c(2, 1, 0))    # axis label at 2 rows distance, tick labels at 1 row

# occurrences
plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border = "grey",
     xlim = c(min(lon), max(lon)), ylim = c(min(lat), max(lat)), main="Observations")
points(as.numeric(distrib[,"long"]), as.numeric(distrib[,"lat"]),
       pch=".", col=terrain.colors(8)[1])

# mask
plot(newmap, col="cornsilk", bg="lightblue1", lwd=0.05, border = "grey",
     xlim = c(min(lon), max(lon)), ylim = c(min(lat), max(lat)),
     main="Summer distribution")
```

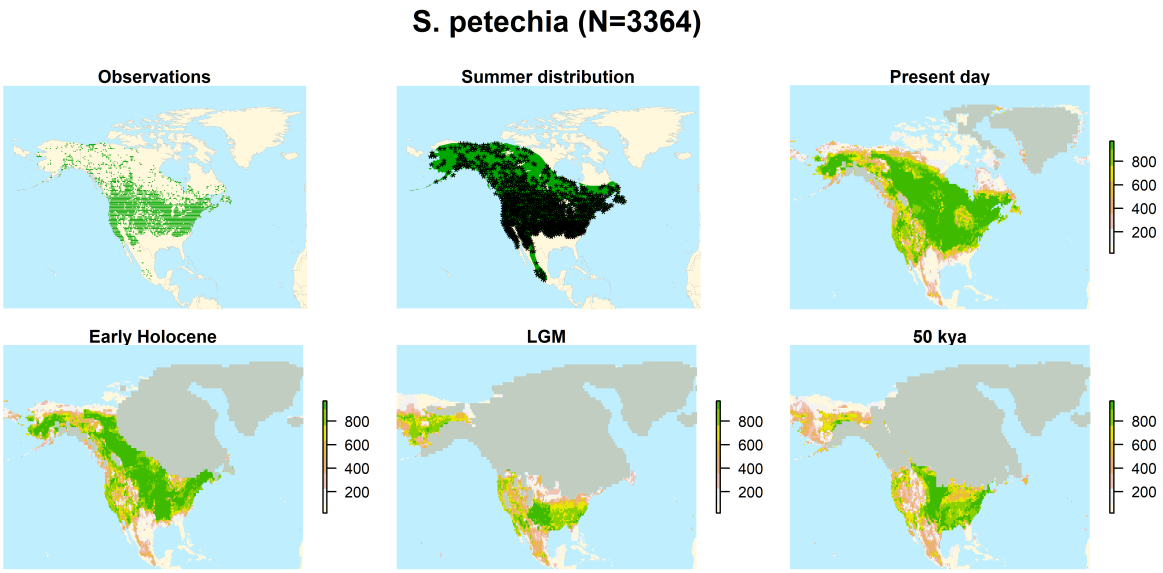


Figure 15: Final plot: original dataset, thinned dataset, summer distribution, and projection in 4 different periods of time based on the mean of the ensemble.

```
plot(myspecies, add=TRUE, col=terrain.colors(8)[1], border=NA)
points(as.numeric(t_dist[,"long"]), as.numeric(t_dist[,"lat"]), pch="*", col="black")

# Ensemble projection modern day
plot(pr,main="Present day", axes=FALSE, box=FALSE, colNA="lightblue1",
     col=c("cornsilk",terrain.colors(8)[8:2]))
plot(r, add=TRUE, col=c(NA,"honeydew3"), border=NA, useRaster=F, legend=FALSE)

# Ensemble projection Holocene
plot(HOL,main="Early Holocene", axes=FALSE, box=FALSE, colNA="lightblue1",
     col=c("cornsilk",terrain.colors(8)[8:2]))
plot(r12, add=TRUE, col=c(NA,"honeydew3"), border=NA, useRaster=F, legend=FALSE)

# Ensemble projection LGM
plot(LGM,main="LGM", axes=FALSE, box=FALSE, colNA="lightblue1",
     col=c("cornsilk",terrain.colors(8)[8:2]))
plot(r21, add=TRUE, col=c(NA,"honeydew3"), border=NA, useRaster=F, legend=FALSE)

# Ensemble projection past
plot(k50,main="50 kya", axes=FALSE, box=FALSE, colNA="lightblue1",
     col=c("cornsilk",terrain.colors(8)[8:2]))
plot(r50, add=TRUE, col=c(NA,"honeydew3"), border=NA, useRaster=F, legend=FALSE)

# title
title(main=paste("S. petechia (N=",dim(distrib)[1],")",sep=""),cex.main= 2,outer = TRUE, line = 2)

dev.off()
```

Save output(s) as netcdf file

The following code saves the output as netcdf files.

```
# read file for the present and create rasters for stats of interest
filelist <- list.files(paste(sp,"/proj_Proj_NAmer/", sep="")) #list of files in directory
enslist <- filelist[grep("ensemble", filelist)]  # selecting consensus projections
enspbin <- stack(paste(sp,"/proj_Proj_NAmer/", enslist[2], sep=""))

# create raster for each statistic
mea <- raster(enspbin,c(1))
med <- raster(enspbin,c(2))
ca <- raster(enspbin,c(3))
wm <- raster(enspbin,c(4))

# do the same for all periods, stack rasters and save as netcdf file
for (i in ref){

  filelist <- list.files(paste(sp,"/proj_",i,"_NAmer/", sep="")) #list of files in directory
  enslist <- filelist[grep("ensemble", filelist)]  # selecting consensus projections
  enspbin <- stack(paste(sp,"/proj_",i,"_NAmer/", enslist[2], sep=""))

  mea <- stack(mea, raster(enspbin,c(1)))
  med <- stack(med, raster(enspbin,c(2)))
  ca <- stack(ca, raster(enspbin,c(3)))
  wm <- stack(wm, raster(enspbin,c(4)))
}

  writeRaster(mea, paste("Yellow_warbler_",ens.methods[1],".nc", sep=""),
              overwrite=TRUE, format="CDF", varname=ens.methods[1],
              longname=paste("Probability (scale: 0-1000) of Yellow warbler presence from SDM analyses"
                             ens.methods[1], sep=" - "),
              xname="Longitude", yname="Latitude", zname="Time (kyrs ago)")

  writeRaster(med, paste("Yellow_warbler_",ens.methods[2],".nc", sep=""),
              overwrite=TRUE, format="CDF", varname=ens.methods[2],
              longname=paste("Probability (scale: 0-1000) of Yellow warbler presence from SDM analyses"
                             ens.methods[2], sep=" - "),
              xname="Longitude", yname="Latitude", zname="Time (kyrs ago)")

  writeRaster(ca, paste("Yellow_warbler_",ens.methods[3],".nc", sep=""),
              overwrite=TRUE, format="CDF", varname=ens.methods[3],
              longname=paste("Probability (scale: 0-1000) of Yellow warbler presence from SDM analyses"
                             ens.methods[3], sep=" - "),
              xname="Longitude", yname="Latitude", zname="Time (kyrs ago)")

  writeRaster(wm, paste("Yellow_warbler_",ens.methods[4],".nc", sep=""),
              overwrite=TRUE, format="CDF", varname=ens.methods[4],
              longname=paste("Probability (scale: 0-1000) of Yellow warbler presence from SDM analyses"
                             ens.methods[4], sep=" - "),
              xname="Longitude", yname="Latitude", zname="Time (kyrs ago)")


#To check the ncdf files
#ncin <- nc_open("Yellow_warbler_Mean.nc")
```

```
#ncin
#nc_close(ncin)
```