1 **Title:** mtDNAcombine: tools to combine sequences from multiple studies

2 **Authors:** Eleanor F. Miller* [a], Andrea Manica [a]

3 **Affiliations:**

4 [a] Department of Zoology, University of Cambridge, Downing street, Cambridge, CB2 3EJ, UK.

5 * Corresponding Author: Eleanor F. Miller, Department of Zoology, University of Cambridge, Downing

6 street, Cambridge, CB2 3EJ, UK. Email: em618@cam.ac.uk

7 **ORCID:**

8 EFM: 0000-0002-3213-5714

9 AM: 0000-0003-1895-450X

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 we present the mtDNAcombine 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.

Key words: demographic history, R package, mitochondrial DNA, public datasets, Bayesian skyline plots

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 [1–3]. 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 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 [4], lack of recombination [5] and uniparental mode of inheritance [6] make it more sensitive to capture changes in population size than slower evolving nuclear genes [7] (Fig. 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) [8], 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.

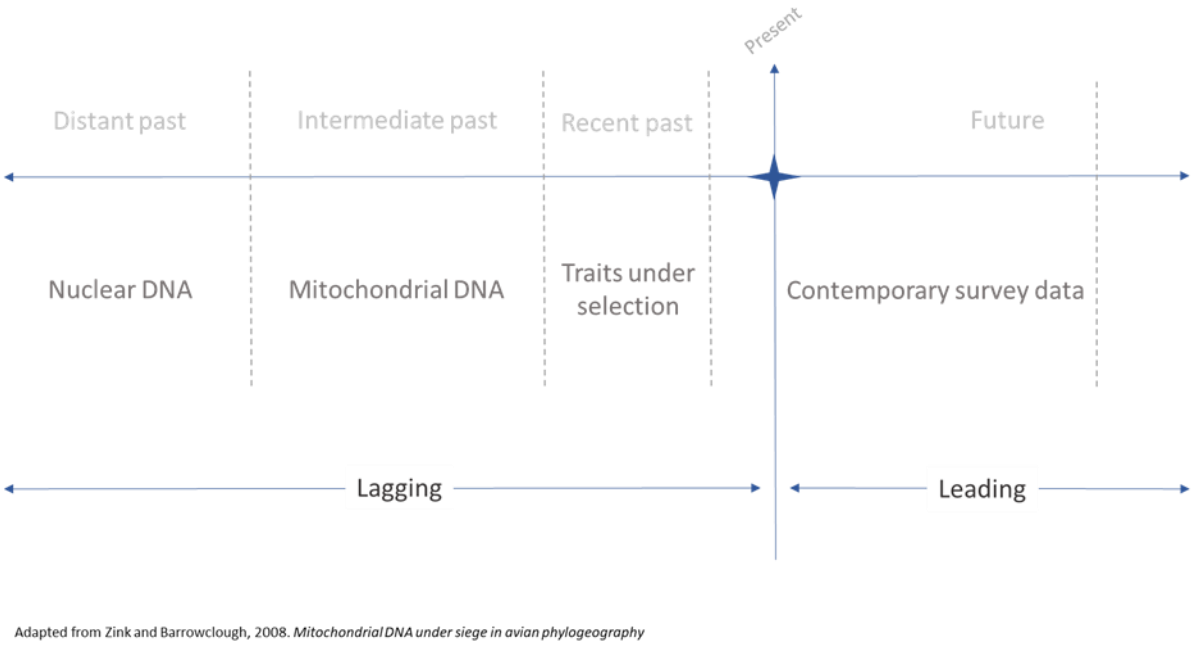

Figure 1. Utility of different loci for reconstructing different periods of population history.

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) [9], the availability of WGS data is rapidly increasing. Using a single, high quality, diploid genome sequence, the pairwise sequentially Markovian coalescent (PSMC) method [10] 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 [11]. 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 [11]. 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 [12]. 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 [13], 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 the current paper, we consider the practicalities and problems faced by a meta-analysis of publicly available data and present the mtDNAcombine package. The mtDNAcombine 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 (Fig. 2.).

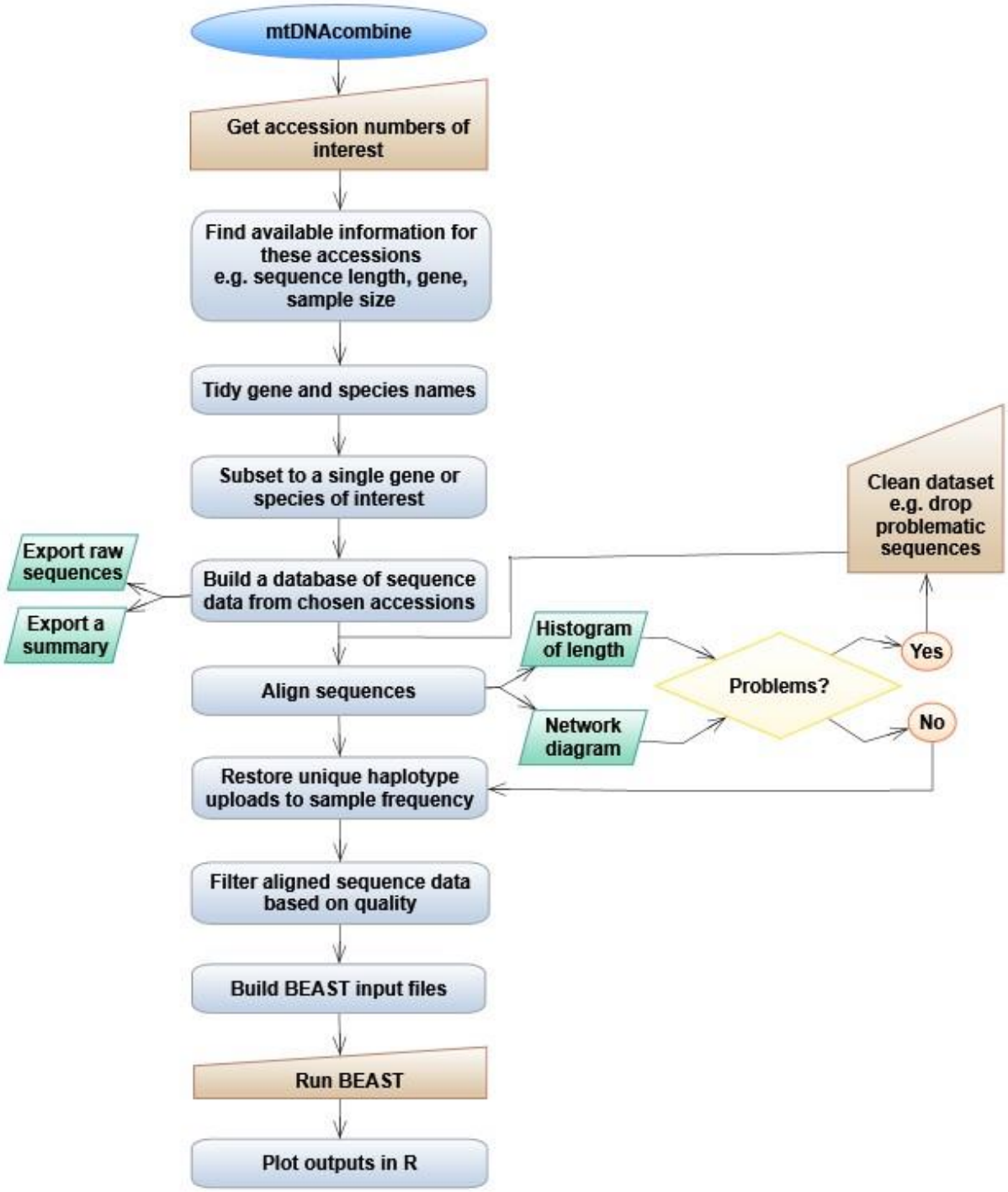

Figure 2. Flow diagram of mtDNAcombine pipeline showing decisions and steps supported by the package.

Methods:

Data preparation

Raw data – Step one is to search annotated DNA databases to determine how many data sets are available. We 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 mtDNAcombine, we import information (e.g. title of associated paper and sequence length) about relevant accessions into a dataframe with the 'build_GB_dataframe' function. We then proceed to explore and clean up this information to make it comparable across studies, and thus allow us to merge data for the any given species and create comparable datasets for multiple species.

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 mtDNAcombine 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 [14] 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 include an underscore ("_") in their accession number, while simple repository accessions never have this character. Our code ('load_accession_list' 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 [15], MUSCLE [16,17] and MAFFT [18]. In mtDNAcombine, we chose to use ClustalW [19], implemented through the R package 'msa'. [20]. Though BEAST can handle missing / ambiguous bases [21], we consider it best to use alignments without gaps or ambiguities. Whilst some insertions or deletions may be

127 genuine, when working with sequences from multiple sources, the data are likely to have been
128 sequenced with different techniques to varying standards. Inclusion of basic sequencing errors could
129 drive miscalculations in later analyses and the volume or type of errors will not be consistent across
130 all studies, nor across all taxa. We therefore recommend that, to ensure consistent sequence quality,
131 all sites with ambiguities, insertions, deletions and missing data should be removed. This is done
132 automatically within the 'align_and_summarise' function in mtDNAcombine.

133

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

142

143 **Sequence length** - For any group of studies there will be numerous reasons the samples were original
144 collected and sequenced. Each project will have had, among other things, a different budget, time
145 constraints, target area of the mitochondrial genome, and available sequencing technology, meaning
146 that different lengths of the genome / target gene will have been sequenced. In some instances,
147 only very short sections of the gene of interest will have been sequenced. If the number of base
148 pairs (bp) is too low, the sample is unlikely to hold enough information to be informative for
149 population demographic reconstruction. The 'align_and_summarise' function will drop individual
150 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 mtDNAcombine produces is a histogram showing the original variation in sequence length as well as the length of the trimmed, maximum overlap, dataset (Fig. 3**,** 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, mtDNAcombine allows the user to go back, review, and revise the input data if they want.

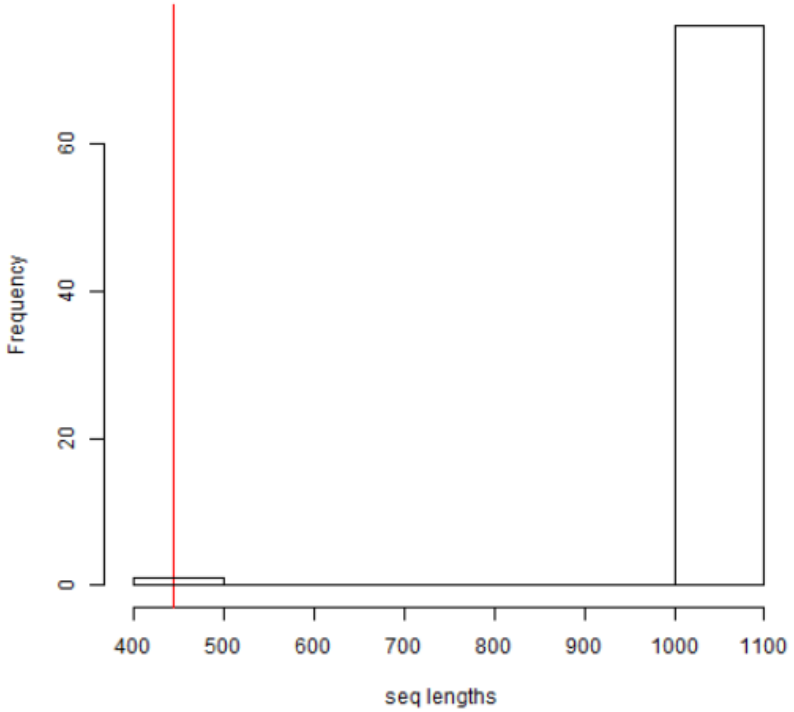

Figure 3. Example of diagnostic plot for sequence trimming in the 'align_and_summarise' function. Histogram shows that, in order to trim all sequences to the maximum overlapping length (red line), the majority of samples have had to be heavily cropped.

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 [22]. 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 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 [23–26]. BSP analysis, like other coalescent methods, is founded on the Wright-Fisher model and hence assumes panmixia [27]. This assumption is violated by population sub-structure [23,28], 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 [23]. 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, we draw network diagrams within R using the package 'pegas' [29]. These network diagrams are one of the diagnostic plots created by the 'align_and_summarise' function (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 – We 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 [30], amplification of integrated nuclear copies, incorrect accession codes, or even result from poor-quality sequencing. We feel that the benefits of including these outliers in case they are genuine are far outweighed by the risk that they distort the process of inference. We 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. We 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.

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 [21] so we 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 [31], 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 [21]. Bearing this in mind, we 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 mtDNAcombine package function 'setup_basic_xml' utilises the 'babette' package [32] to build basic XML files form the data set 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, we encourage use of the BEAGLE library [33] 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. [22,23] and by those who designed and built the software [34–38]. 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 [39] 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' [32], we think 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 we do in mtDNAcombine.

Plotting profiles in R – BSPs can be drawn using the programme Tracer [39]. However, for more flexibility, and to facilitate exploration of the profiles in greater detail, we chose to visualise the reconstructed profiles in R. Within the mtDNAcombine package vignette, we 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 [22,40], 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 [22,23]. 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 (Fig. 4.). 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.

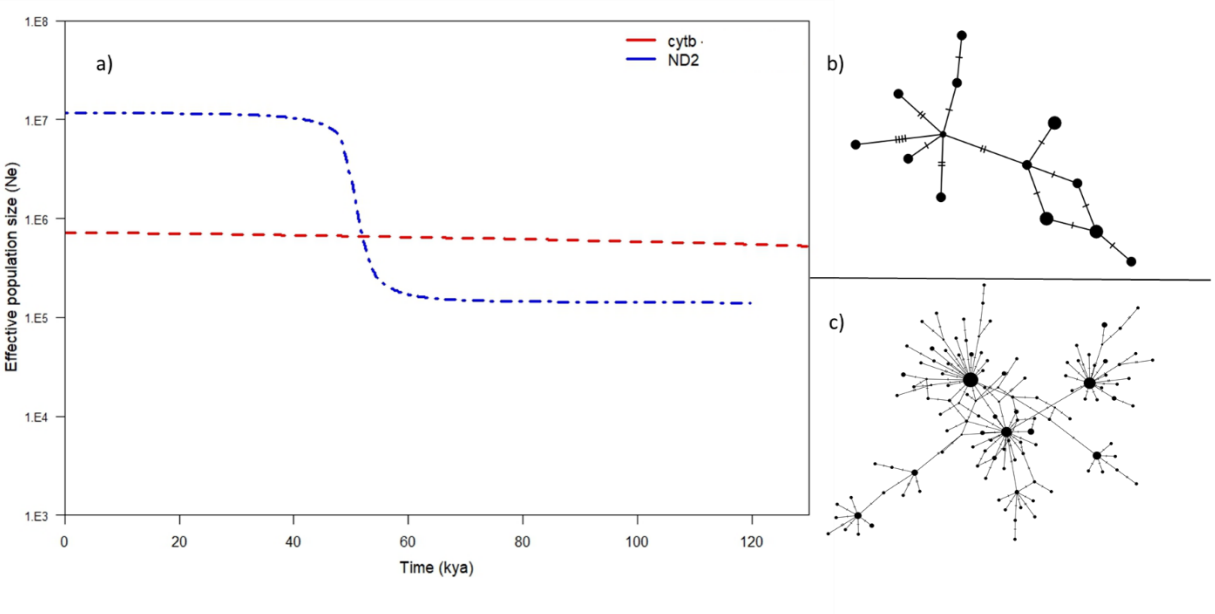

Figure 4. Comparison of two dissimilar BSP profiles drawn from different mtDNA datasets of the common rosefinch. a) Red line is median value for cytb BSP profile, blue line is median value for ND2 BSP profile. The cytb dataset includes 15 samples, ND2 dataset 190 samples. The varying levels of information 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. We 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, we 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.

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. [41–43]. 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 we present the mtDNAcombine package, providing a pipeline to streamline the process of downloading, curating and analysing mitochondrial sequence data (Fig. 2). 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 we 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. We 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. We hope that our discussion, whilst highlighting common pitfalls, provides solutions and suggestions to guide the process of compiling data sets from online databases.

Funding:

E.F.M was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) Doctoral Training Partnerships program (grant code: BB/M011194/1).

Data Availability:

mtDNAcombine source code and full vignette can be found at: https://github.com/EvolEcolGroup/mtDNAcombine . The data used as examples in this paper were derived from publicly available data in GenBank (https://www.ncbi.nlm.nih.gov/genbank/). A sample dataset is available within the R package.

Author contributions:

E.F.M. performed the analyses, prepared all figures and wrote the manuscript. All authors conceived the idea and design, as well as reviewing the manuscript.

Additional information:

The authors declare no competing financial interests.

1. Nadachowski A. 1989 Origin and History of the Present Rodent Fauna in Poland Based on Fossil Evidence. *Acta Theriol. (Warsz).* **34**.

2. Sommer RS, Benecke N. 2005 The recolonization of Europe by brown bears Ursus arctos Linnaeus, 1758 after the Last Glacial Maximum. *Mamm. Rev.* **35**, 156–164. (doi:10.1111/j.1365-2907.2005.00063.x)

3. Sommer R, Benecke N. 2005 Late-Pleistocene and early Holocene history of the canid fauna of Europe (Canidae). *Mamm. Biol.* **70**, 227–241. (doi:10.1016/j.mambio.2004.12.001)

4. Brown WM, George M, Wilson AC. 1979 Rapid evolution of animal mitochondrial DNA. *Annu. Rev. Ecol. Syst.* **18**, 269–292. (doi:10.1146/annurev.es.18.110187.001413)

5. Elson JL, Andrews RM, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 2002 Analysis of European mtDNAs for Recombination. *Am. J. Hum. Genet.* **68**, 145–153. (doi:10.1086/316938)

6. Giles RE, Blanc H, Cann HM, Wallace DC. 1980 Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **77**, 6715–9.

7. Zink RM, Barrowclough GF. 2008 Mitochondrial DNA under siege in avian phylogeography. *Mol. Ecol.* **17**, 2107–2121. (doi:10.1111/j.1365-294X.2008.03737.x)

8. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192. (doi:10.1093/molbev/msi103)

9. Zhang G. 2015 Genomics: Bird sequencing project takes off. *Nature* **522**, 34. (doi:10.1038/522034d)

10. Li H, Durbin R. 2011 Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496. (doi:10.1038/nature10231)

11. Schiffels S, Durbin R. 2014 Inferring human population size and separation history from multiple genome sequences. *Nat. Publ. Gr.* **46**, 919–925. (doi:10.1038/ng.3015)

12. Garrick RC *et al.* 2015 The evolution of phylogeographic data sets. *Mol. Ecol.* **24**, 1164–1171. (doi:10.1111/mec.13108)

13. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013 GenBank. *Nucleic Acids Res.* **41**, 36–42. (doi:10.1093/nar/gks1195)

14. O'Leary NA *et al.* 2016 Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745. (doi:10.1093/nar/gkv1189)

15. Notredame C, Higgins DG, Heringa J. 2000 T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217. (doi:10.1006/jmbi.2000.4042)

16. Edgar RC. 2004 MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797. (doi:10.1093/nar/gkh340)

17. Edgar RC. 2004 MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 1–19. (doi:10.1186/1471-2105-5-113)

18. Katoh K, Misawa K, Kuma K, Miyata T. 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066. (doi:10.1093/nar/gkf436)

19. Larkin MA *et al.* 2007 Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948. (doi:10.1093/bioinformatics/btm404)

20. Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. 2015 Msa: An R package for multiple sequence alignment. *Bioinformatics* **31**, 3997–3999.

(doi:10.1093/bioinformatics/btv494)

21. Drummond AJ, Bouckaert RR. 2015 *Bayesian evolutionary analysis with BEAST*. Cambridge University Press.

22. Grant WS. 2015 Problems and cautions with sequence mismatch analysis and Bayesian skyline plots to infer historical demography. *J. Hered.* **106**, 333–346. (doi:10.1093/jhered/esv020)

23. Heller R, Chikhi L, Siegismund HR. 2013 The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS One* **8**, e62992. (doi:10.1371/journal.pone.0062992)

24. Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. 2010 The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186**, 983–995. (doi:10.1534/genetics.110.118661)

25. Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**, 205–216. (doi:10.1534/genetics.108.094904)

26. Pannell JR. 2003 Coalescence in a Metapopulation with Recurrent Local Extinction and Recolonzation. **57**, 949–961.

27. Pybus OG, Rambaut A, Harvey PH. 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–37.

28. Ho SYW, Shapiro B. 2011 Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.* **11**, 423–434. (doi:10.1111/j.1755-0998.2011.02988.x)

386 29. Paradis E. 2010 Pegas: An R package for population genetics with an integrated-modular
387 approach. *Bioinformatics* **26**, 419–420. (doi:10.1093/bioinformatics/btp696)

388 30. McMahon MM, Sanderson MJ. 2006 Phylogenetic supermatrix analysis of GenBank sequences
389 from 2228 papilionoid legumes. *Syst. Biol.* **55**, 818–836. (doi:10.1080/10635150600999150)

390 31. Nabholz B, Lanfear R, Fuchs J. 2016 Body mass-corrected molecular rate for bird
391 mitochondrial DNA. *Mol. Ecol.* **25**, 4438–4449. (doi:10.1111/mec.13780)

392 32. Bilderbeek RJC, Etienne RS. 2018 babette : BEAUti 2, BEAST2 and Tracer for R. *Methods Ecol.*
393 *Evol.* **9**, 2034–2040. (doi:10.1111/2041-210X.13032)

394 33. Ayres DL *et al.* 2012 BEAGLE: An Application Programming Interface and High-Performance
395 Computing Library for Statistical Phylogenetics. *Syst. Biol.* **61**, 170–173.
396 (doi:10.1093/sysbio/syr100)

397 34. Atkinson QD, Gray RD, Drummond AJ. 2007 mtDNA Variation Predicts Population Size in
398 Humans and Reveals a Major Southern Asian Chapter in Human Prehistory.
399 (doi:10.1093/molbev/msm277)

400 35. Drummond AJ, Rambaut A. 2007 BEAST : Bayesian evolutionary analysis by sampling trees. **8**,
401 1–8. (doi:10.1186/1471-2148-7-214)

402 36. Heled J, Drummond AJ. 2008 Bayesian inference of population size history from multiple loci.
403 **15**, 1–15. (doi:10.1186/1471-2148-8-289)

404 37. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012 Bayesian phylogenetics with BEAUti and
405 the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973. (doi:10.1093/molbev/mss075)

406 38. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A,
407 Drummond AJ. 2014 BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS*

Comput. Biol. **10**, 1–6. (doi:10.1371/journal.pcbi.1003537)

39. Rambaut A, Suchard MA, Xie D, Drummond AJ. In press. Tracer v1.6.

40. Grant WS, Liu M, Gao T, Yanagimoto T. 2012 Limits of Bayesian skyline plot analysis of mtDNA sequences to infer historical demographies in Pacific herring (and other species). *Mol. Phylogenet. Evol.* **65**, 203–212. (doi:10.1016/j.ympev.2012.06.006)

41. Hope AG, Ho SYW, Malaney JL, Cook JA, Talbot SL. 2014 Accounting for rate variation among lineages in comparative demographic analyses. *Evolution (N. Y).* **68**, 2689–2700. (doi:10.1111/evo.12469)

42. Antonelli A *et al.* 2017 Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of Taxa. *Syst. Biol.* **66**, 153–166. (doi:10.1093/sysbio/syw066)

43. Smith SA, Beaulieu JM, Donoghue MJ. 2009 Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol. Biol.* **9**, 37. (doi:10.1186/1471-2148-9-37)