Detailed description of the research program

1 Scientific background

The primary objective of research in evolution is to understand how biodiversity is generated in nature. Speciation, the process in which two populations diverge to form two distinct species, has thus been the focus of extensive research (e.g., [12, 13, 11, 39]). Up until recently, research in this field has been severely limited by sparse genomic data, but the advent of high throughput DNA sequencing (HTS) is looking to change that by allowing researchers to obtain high quality sequence data at low cost. As a result, we are seeing an increasing number of speciation studies that sequence the genomes of multiple individuals from closely related species. With the data barrier practically lifted, the main challenge lies in developing computational and statistical methods that will aid evolutionary biologists in the study of recently diverged species [49].

This central challenge has been addressed during the past decade by developing sophisticated demography inference methods, which estimate divergence times, ancestral effective population sizes, and rates of gene flow by analyzing patterns of sequence variation in a collection of sampled individuals. One particularly fruitful approach for demography inference has been to explicitly represent the genealogical relationships between the sampled individuals at numerous (approximately) unlinked loci, and use these local genealogies as hidden variables in a Markov chain Monte Carlo (MCMC) sampling algorithm to produce Bayesian estimates of the demographic parameters. This approach was initially developed in two separate methods: *IM* [24, 23], for analysis of two populations with divergence and gene flow, and *MCMCcoal* [46, 5], for analysis of multiple populations with tree-like divergence and no gene flow. Recently, we developed the Generalized Phylogenetic Coalescent Sampler (G-PhoCS), which generalizes both methods by assuming a multi-population demographic model with gene flow. G-PhoCS is also implemented to be significantly more efficient than its predecessors ($\sim$10-fold [20]), enabling it to analyze tens of thousands of loci sequenced from up to 20 individuals. G-PhoCS was originally developed to infer ancient human demography [20], and subsequently applied to canids [19] and primates [9, 35]. The efficiency and flexible model of G-PhoCS make it particularly suitable for analysis of HTS data sets that are emerging in speciation study.

While there exist many other methods for demography inference, the genealogy sampling approach is appealing for several key reasons. Unlike popular methods for population structure inference (e.g., *STRUCTURE* [43, 1]), genealogy samplers assume an explicit demographic model, and unlike methods based on approximate Bayesian computation (ABC; e.g. [2, 14, 48]), they rely on an explicit likelihood function and do not require prior assumptions on which patterns in the sequence data would be informative about the demographic history. The small number of linked SNPs in each locus provide valuable (yet weak) information about the structure of the local genealogy, which cannot be effectively utilized by methods that analyze unlinked SNPs (e.g., ∂a∂i [21]). Finally, the use of short loci makes these genealogy sampling methods particularly suitable for reduced representation sequence data, such as ddRADtag, which is heavily used in the study of non-model organisms.

2 Research Objectives and Expected Significance

The primary objective of the proposed study is to develop novel computational methods for studying recently diverged species, and to apply these methods to large-scale genome sequence data to provide new insights on the process of speciation. Progress will be made in three parallel tracks: (1) theory and methods development, (2) analysis of data from two avian species groups generated by our collaborators, and (3) development of public software tools. Synergy will be maintained between these three tracks by using the sequence data to guide the methods development and using our experience in data analysis to construct user-friendly tools geared toward a wide community of evolutionary biologists. Through this strategy we wish to unlock the great potential of HTS data in addressing fundamental questions having to do with speciation.

Objective 1: computational methods for Bayesian hypothesis testing

When conducting evolutionary inference from recently diverged species, we are often interested in structural features of that history, such as the topology of the species tree, admixture events, and mode of gene flow (continuous versus secondary contact). The problem is that demography inference methods are designed to infer parameters in an assumed demographic model (i.e., divergence times, effective population sizes, and migration rates), but they provide little information about the fit of the assumed model to the data. Thus researchers often end up making inferences about structural features of the model through ad-hoc interpretation of the estimated parameters (see, e.g., our study about the origin of domestic dogs in [19]). We will develop methods for testing and ranking different demographic models based on their fit to the data. These methods will make use of the genealogy sampling technique implemented in demography inference methods such as G-PhoCS or *IM*. We will ensure that these methods are sufficiently accurate and powerful to make useful inferences about the structure of divergence of closely related species.

Expected significance: The newly developed methods will solve an important open problem with existing demography inference methods, and will allow us to conduct a robust statistical evaluation of alternative hypotheses regarding speciation.

Objective 2: gene flow during speciation in two avian case studies

As case studies for recent species radiation we will use two groups of bird species: the capuchino seedeaters within the genus *Sporophila* [6] and the *Setophaga* (previously *Dendroica*) wood warblers [34]. These two groups have been extensively studied by our collaborators from the Lab of Ornithology at Cornell University (see attached letter of collaboration from Prof. Irby Lovette). Genome sequence data obtained for these two groups will be used as a benchmarks for methods development, and as means for producing new insights about the process of speciation. We are particularly interested in using these case studies to examine the role of gene flow during species divergence. Species in these two groups were thus carefully selected to cover a variety of scenarios (times of divergence and allopatry) and ensure that our findings could be generalized to other cases.

Expected significance: Analysis of sequence data from these two avian groups will provide an important proof of concept for our the newly developed methods, and will produce generalizable insights about speciation.

Objective 3: produce user friendly public software tools
The computational methods developed in this study will be implemented in open source and will be made publicly available. The published software will be supplemented with user-friendly interface modules that facilitate examination and evaluation of multiple evolutionary hypotheses. These modules will guide the user in setting up an analysis and will assist in visualizing and interpreting the results.

Expected significance: Constructing a carefully designed interface for our software will answer a well acknowledged need in evolutionary biology for powerful, yet user-friendly, inference tools. This will enable a broad community of researchers to realize the full potential of HTS data in speciation studies.

3 Detailed Description of the Proposed Research

Achieving the objectives listed above requires developing new theory and methods alongside extensive data analysis and software development. A detailed outline of our plans in these three parallel tracks is described in Sections 3.1–3.3 below. We follow by summarizing our preliminary results (Section 3.4) and providing additional details about our research plan (Sections 3.5 and 3.6).

3.1 Research design and methods – computational methods

Bayesian demography inference by MCMC

Demography inference methods typically take in sequence data ($\mathcal{X}$) from a collection of individuals from closely related populations and a parameterized demographic model (M), and they infer values of parameters Θ in that model. Bayesian methods achieve this by assuming some prior distribution on the model parameters ($P(\Theta|M)$) and sampling parameter values from an approximate posterior distribution ($P(\Theta|M, \mathcal{X})$. Because the joint probability distribution $P(\mathcal{X}, \Theta|M)$ cannot be efficiently computed, this task is often done by introducing hidden variables $\mathcal{Z}$ to the model, such that the probability $P(\mathcal{X}, \mathcal{Z}, \Theta|M)$ can be efficiently and accurately computed, and employing an MCMC sampling algorithm for $\mathcal{Z}$ and Θ. Sampling by MCMC guarantees that $(\Theta, \mathcal{Z})$ will be sampled from a probability distribution approximating the posterior–$P(\Theta, \mathcal{Z}|\mathcal{X}, M)$ [36, 22]. From this distribution one can extract approximate posterior means and credible intervals for all demographic parameters.

For instance, the demographic model M assumed by G-PhoCS is defined using a population phylogeny augmented by horizontal edges called *migration bands*, and the demographic parameters Θ consist of divergence times (τ) associated with internal nodes of the phylogeny, effective population sizes (θ) associated with its branches, and migration rates (m) associated with migration bands (Fig. 1A). A product of Gamma distributions is used as a prior $P(\Theta|M)$ for model parameters. The data examined by G-PhoCS ($\mathcal{X}$) are a collection of multiple sequence alignments (X_l) at L short loci assumed to be genetically unlinked and neutrally evolving (Fig. 1A). The hidden variables of the model ($\mathcal{Z}$) are a collection of local genealogies (or gene trees), $\mathcal{Z} = \{G_l\}_{l=1}^{L}$, one for each locus. Because the loci are assumed to be unlinked, the likelihood is given by a product across loci, and the contribution of each locus to the likelihood consists of a genealogy prior, $P(G_l|M)$, based on coalescent theory [26, 3, 38], and the data likelihood, $P(X_l|G_l)$, defined using a DNA substitution model (e.g. [28, 29, 18]).

$$P(\mathcal{X}, \mathcal{Z}, \Theta|M) \;=\; P(\Theta|M) \prod_{l=1}^{L} P(G_l|M) P(X_l|G_l)\,. \tag{1}$$

Relative Bayes factors

Evolutionary inference is based on our ability to measure the statistical fit of different models to data. For instance, demographic inference makes certain assumptions about the demographic model M (e.g., structure of the population phylogeny, mode of gene flow), and then infers parameters under that model. However, the inferred parameter values do not provide much information about the fit of the assumed model to data. Model fit can be measured by the likelihood $P(\mathcal{X}|M)$, which is also known as the *Bayes factor*. Bayes factors are the key for conducting Bayesian hypothesis testing, but they are notoriously difficult to estimate for complex models [37]. **We propose to address this problem by developing methods for model comparison based on *relative Bayes factors* with respect to some reference model $\mathbf{M_0}$: $\mathbf{BF(M_0 : M) = P(X|M_0)/P(X|M)}$.**

We outline our proposed approach below. Let M be an arbitrary demographic model, which in our case consists of a population phylogeny, migration bands, and prior distribution on model parameters. Let $\mathcal{Z}\Theta$ denote the aggregation of the model parameters (Θ) and the hidden variables of the model ($\mathcal{Z}$; local genealogies in our case). A model M_0 is considered a generalization of M if it obeys the following two conditions: (1) there is a mapping between parameters in M onto parameters in M_0, and (2) if $P(\mathcal{Z}\Theta|M)$ is greater than 0, then so is $P(\mathcal{Z}\Theta|M_0)$. If these conditions are met, then the relative Bayes factor $BF(M_0 : M)$ given data $\mathcal{X}$ can be expressed as the following expectation under the posterior distribution $P(\mathcal{Z}\Theta|\mathcal{X}, M)$:

$$\begin{aligned} BF(M : M_0) \;&=\; \frac{P(\mathcal{X}|M_0)}{P(\mathcal{X}|M)} \;=\; \int \frac{P(\mathcal{Z}\Theta, \mathcal{X}|M_0)}{P(\mathcal{X}|M)} d\mathcal{Z}\Theta \;=\; \int \frac{P(\mathcal{Z}\Theta, \mathcal{X}|M_0)}{P(\mathcal{Z}\Theta, \mathcal{X}|M)} P(\mathcal{Z}\Theta|\mathcal{X}, M) d\mathcal{Z}\Theta \\ &=\; \int \frac{P(\mathcal{Z}\Theta|M_0)}{P(\mathcal{Z}\Theta|M)} P(\mathcal{Z}\Theta|\mathcal{X}, M) d\mathcal{Z}\Theta \;=\; \mathbb{E}_{\mathcal{Z}\Theta|\mathcal{X},M}\left(\frac{P(\mathcal{Z}\Theta|M_0)}{P(\mathcal{Z}\Theta|M)}\right)\,. \end{aligned} \tag{2}$$

Implementation

A direct consequence of the expression in (2) is that the relative Bayes factor can be approximated by generating N samples $\{(\mathcal{Z}\Theta)_n\}_{n=1}^{N}$ of hidden variables and model parameters using MCMC sampling from an approximate posterior distribution, and then taking the mean of the ratios $\frac{P((\mathcal{Z}\Theta)_n|M_0)}{P((\mathcal{Z}\Theta)_n|M)}$. A major benefit of this approach is that it fits naturally within the existing framework of demography inference methods based on MCMC sampling. **We will implement this scheme for approximating relative Bayes factors in the source code of G-PhoCS**. In each sampling iteration, in addition to the sampled values of model parameters (Θ), the algorithm will also output sufficient statistics for computing $P(\mathcal{Z}\Theta|M_0)$ for every generalization M_0 of M. Defining an adequate set of sufficient statistics will make this method both **generalizable and scalable**. Generalization is crucial because it enables estimation of Bayes factors of M relative to various reference models using a single run of the sampler. Scalability will be achieved by defining a set of statistics that can be aggregated across the L loci, so that the number of additional recorded values does not increase with the number of loci.

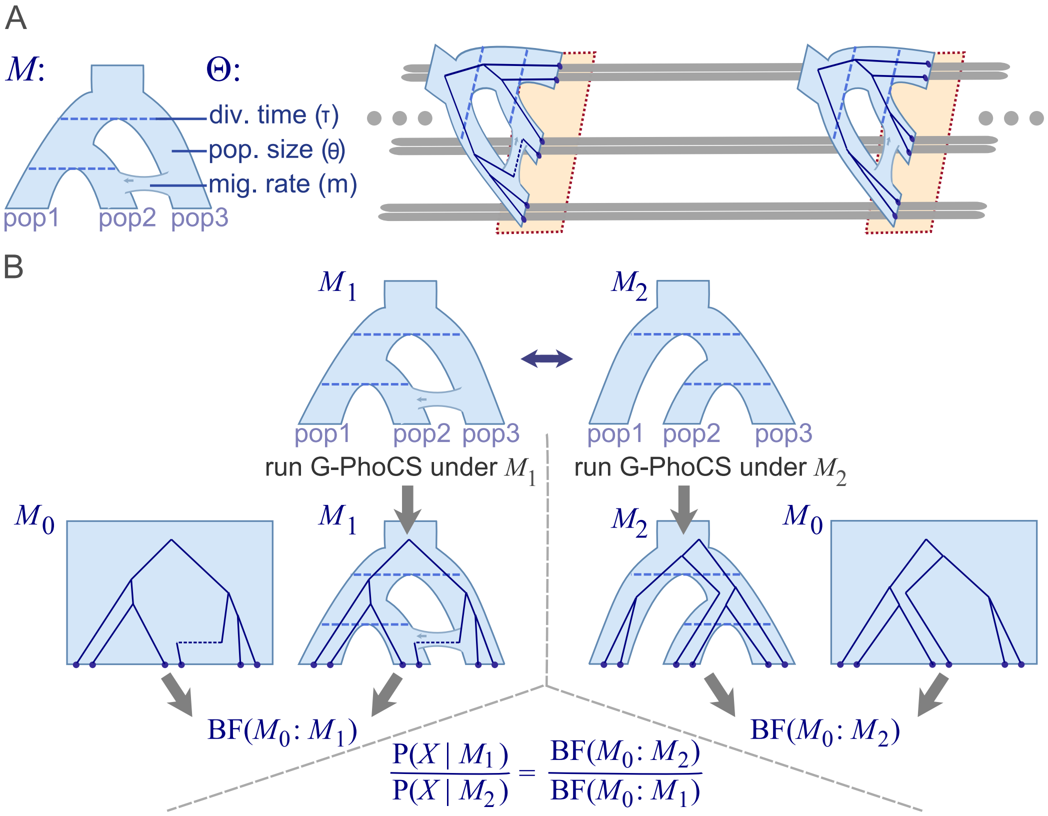

Figure 1. Bayesian model comparison with G-PhoCS. **(A)** G-PhoCS infers demographic parameters (Θ) along a population phylogeny augmented with migration bands (M; left). The input consists of the assumed model (M) and multiple sequence alignments $\mathcal{X} = \{X_l\}_{l=1}^{L}$ for the sequenced individuals at short inter-spaced genomic loci (right). Values of the demographic parameters (Θ) are sampled together with local genealogies $\mathcal{Z} = \{G_l\}_{l=1}^{L}$ (one for each locus) according to a distribution that approximates the posterior–$P(\Theta, \mathcal{Z}|\mathcal{X}, M)$. **(B)** We propose to develop methods for comparing different evolutionary hypotheses using the genealogy sampler of G-PhoCS. In the illustrated example, one hypothesis (M_1) assumes that pop1 and pop2 are sister populations and there is gene flow from pop3 into pop2, and the alternative hypothesis (M_2) assumes that pop2 and pop3 are sister populations and there is no post-divergence gene flow (top). To compare the two hypotheses, G-PhoCS is run on the input alignments $(\mathcal{X})$ separately with model M_1 and with model M_2. The Bayes factor for a given reference model M_0 (a single panmictic population in this case) relative to M_i, $BF(M_0 : M_1)$, is computed by taking the average of the ratio $P(\mathcal{Z}\Theta|M_0)/P(\mathcal{Z}\Theta|M_i)$ across all genealogies and parameter values $\mathcal{Z}\Theta$ sampled by the MCMC algorithm when run with model M_i (equation (2)). Then, the relative fit of M_1 and M_2 to the data is estimated using the ratio of the relative Bayes factors (bottom). Robust estimates of this relative fit can be obtained by using various models as reference (M_0).

Estimation variance

The proposed method for estimating relative Bayes factors is based on importance sampling (IS), which is a technique used for inferring properties of a probability distribution ($\Theta|\mathcal{X}, M_0$ in this case) by sampling from a related distribution ($\Theta|\mathcal{X}, M$). Importance sampling (IS) is an extremely useful tool for statistical inference, but its fundamental limitation is that it results in estimation variance that tends to grow as the distance between the sampling distributions and the distribution of interest grows. This has been particularly noticed in applications of IS to demography inference [30, 25]. We thus expect that the approximation of $BF(M_0 : M)$ will be more accurate the more similar the reference model M_0 is to M. In this context, relative Bayes factors have a significant advantage over the *harmonic mean* estimator [37], which is based on the following expression (derived similarly to (2)):

$$\frac{1}{P(\mathcal{X}|M)} \;=\; \mathbb{E}_{\mathcal{Z}\Theta|\mathcal{X},M}\left(\frac{1}{P(\mathcal{X}|\mathcal{Z}\Theta, M)}\right) . \tag{3}$$

The harmonic mean is currently the only existing general method to approximate $P(\mathcal{X}|M)$ using MCMC. Its clear advantage is that it does not require defining a reference model, but it is notoriously difficult to apply to real data because it suffers from unbounded estimation variance [37]. Introducing an explicit reference model (M_0) significantly reduces this variance because the ratio $P(\mathcal{Z}\Theta|M_0)/P(\mathcal{Z}\Theta|M)$ used in the computation of $BF(M_0 : M)$ (see equation (2)) is typically much less variable than $1/P(\mathcal{X}|\mathcal{Z}\Theta, M)$. **Moreover, the ability to examine various reference models will enable robust application of this approach for Bayesian hypothesis testing** (see below).

Ranking models based on relative Bayes factors

The typical scenario we would like to address is one where a collection of plausible evolutionary hypotheses needs to be tested or ranked according to their fit to a given sequence data set ($\mathcal{X}$). These evolutionary hypotheses are given in the form of K demographic models $M_1, \ldots, M_K$, which differ in the structure of the population phylogeny, migration bands, or mode of gene flow (continuous vs. burst, constant along genome vs. variable). These models can be ranked based on their approximate likelihoods ($P(\mathcal{X}|M_k)$) by selecting an appropriate reference model M_0 and estimating a relative Bayes factor $BF(M_0 : M_k)$ for each model using K separate Markov chains (K runs of G-PhoCS). The selected reference model M_0 should be a generalization of all K models, but it should also be as similar as possible to all of them, to reduce the estimation variance. This process is schematically demonstrated in Figure 1B with a reference model (M_0), which has a single panmictic population. This reference model generalizes all other models, and is thus the default choice for reference. However, the same comparison could be carried out using a different reference model M_0', which has a single ancestral population which splits at a given time into the three sampled populations. This alternative reference model represents more population structure and is more similar than the panmictic model to both M_1 and M_2. Thus we expect it to result in lower estimation variance, making it a better choice for a reference in this comparison. Hence, various reference models should be considered in each model comparison or ranking to ensure robust results. We will examine the influence of the choice of reference model on the results using simulated data (see below).

Suggesting alternative models

To complement the information provided by relative Bayes factors, which enables the ranking of a given collection of hypotheses, we will develop methods for proposing new alternative models. This will be done by collecting summaries of the sampled genealogies ($\mathcal{Z}$) that indicate possible lack of model fit. Methods that seek evidence of lack of model fit have been successfully used to detect admixture and gene flow by rejecting a pure phylogenetic model [15, 41, 33]. However, these methods make use of simple patterns of mutations (e.g., ABBA-BABA), which limits their use to small numbers of individuals (typically four) and examination of one admixture event at a time. We will develop methods that make explicit use of genealogies sampled from an approximate posterior distribution$P(\mathcal{Z}|\mathcal{X}, M)$. By recording features that deviate from the assumed model in the genealogies sampled by G-PhoCS, we will obtain information on possible faulty model assumptions. For example, we will record the identity of the lineages coalescing first in an ancestral population. The distribution of these identity pairs is expected to be uniform and symmetric across the different populations. Observed asymmetries will be used to indicate unmodeled gene flow. D-statistics [15] achieve this for the simple case of four lineages and one possible admixture event. Using the more general setup in G-PhoCS, with multiple populations and migration bands, and explicit local genealogies, will allow us to extend this test to a larger number of lineages and more complex scenarios of gene flow. We will also use these statistics to measure the influence of gene flow on different loci to help us determine whether gene flow is homogeneous genome-wide, or whether it is particularly prohibited in certain islands of speciation (see below).

Testing on simulated data

We will conduct extensive simulations to validate the developed methods and help bridge between the theory presented here and practice. Simulations will be conducted using the coalescent simulator ms [26], which we have previously used to validate G-PhoCS [20, 19]. We will design simulated data based on the preliminary demographic inference conducted on the two avian radiations, to ensure that the methods are well suited for our main case study (see **Preliminary Analyses**). Using the case studies as a base line, we will simulate a wide range of demographic scenarios to ensure generality of the novel methods. Using simulated data, we intend to examine the influence of the choice of the reference model M_0 on the approximate Bayes factors computed for it. In particular, we will study how refinements of the default panmictic reference model help improve accuracy by reducing estimation variance. Our main source of validation will come from simulating data under a certain model and then using the data and approximate relative Bayes factors to rank the simulated model with respect to other similar models. We will also assess the accuracy of our approximate Bayes factors by comparing them to Bayes factors estimated by more computationally intensive methods, such as thermodynamic integration [32]. Because thermodynamic integration reduces the estimation variance of IS-based estimates at a relatively high computational cost, this type of validation will be applied only to smaller data sets. Another issue we intend to examine using simulated data is the amount of information in the genealogy summaries described above on lack of model fit. We will focus on tests for admixture by simulating data under complex scenarios of gene flow with multiple admixture events and then testing whether different genealogy statistics can be used to reject models that do not contain certain migration bands.

3.2 Research design and methods – avian speciation case studies

Rationale for choice of the avian groups

The data analysis component of this project is devoted to the study of two cases of species radiation in birds, the *Sporophila* capuchino seedeaters and the *Setophaga* wood warblers, which were chosen for several key reasons:

- Both groups present examples of speciation events that took place in a range of 1-10 million years ago [44, 45, 34, 8]. Thus while all events were fairly recent, they provide a sufficiently large range of times to produce results that can be generalized to other cases.
- Despite their shallow divergence, many of these taxa have been established as distinct biological species through field experiments that assessed pre-mating reproductive barriers, such as female choice based on male plumage pattern and song (e.g., [4, 50])
- The geographical ranges of species within each group have different degrees of overlaps. Thus these species will provide information on the effects of gene flow on genomes under different degrees of sympatry and allopatry (Fig. 2).
- The research group of Prof. Irby Lovette, with whom we will be closely collaborating on this analysis, has extensive experience with the biology and phylogenetics of both species groups and has obtained the necessary genetic samples (see **Preliminary results**).

Genome sequence data

The Lovette lab is sequencing the genomes of individuals from these species using state-of-the-art high throughput technology. The sequencing effort is part of a greater study intended to explore phylogenetic relationships in these two clades, and the analysis done by us will largely complement that study. Current sequencing efforts use double digest restriction site associated DNA tags (ddRADtag; [40]) on Illumina sequencing platforms, This approach results in high quality genotyping of densely sampled loci across the genome, and is thus ideally suitable for analysis using the genealogy sampler approach of G-PhoCS, which is designed to analyze interspersed loci (Fig. 1A). The current strategy is to cover up to ten species in each of the two avian groups and sequence roughly five individuals per species. Preliminary data for Sporophila (see Section 3.4) indicates that this approach of sequencing loci covering roughly 1% of the genome provides sufficient statistical power for demography inference. To assess the effect of using ddRADtag sequencing instead of whole-genome resequencing, the Lovette lab will also sequence and assemble a reference genome for each group—one for *Sporophila hypoxantha* and one for *Setophaga virens*—and use these sequences to resequence a small number of individuals in each group. Combining different types of data (ddRADtag and whole genome resequencing) will allow us to avoid the pitfalls of each of them—small number of individuals for resequencing, and lack of spacial information for loci in ddRADtag.

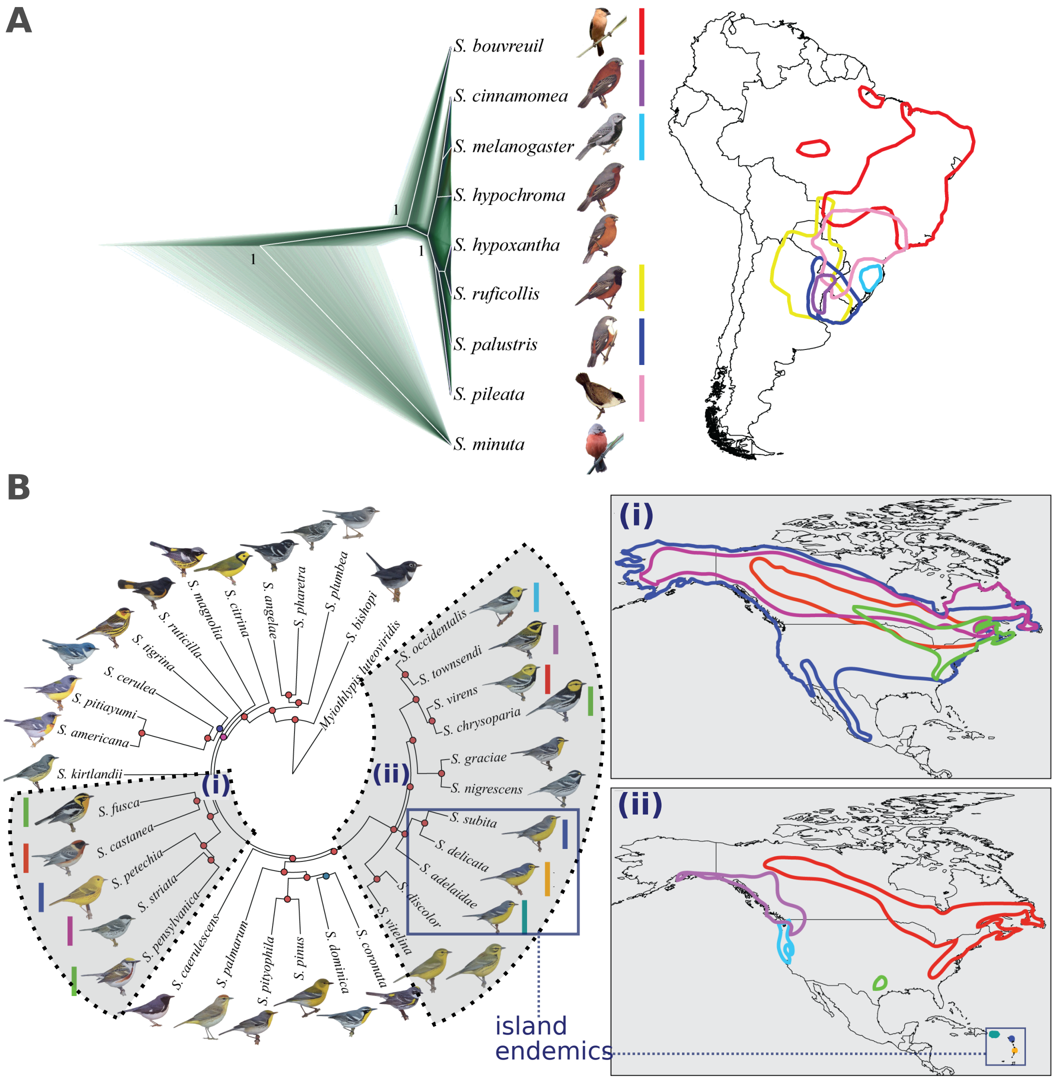

Figure 2. Two avian models. (A) Capuchino seedeaters of the genus *Sporophila*. The estimated phylogeny (left) and geographic ranges (right) of the eight species in our study and a more distant outgroup species, *Sporophila minuta*. The species ranges exhibit different degrees of sympatry/allopatry. The tree was inferred from mitochondrial and nuclear sequence data from 95 individuals. The consensus tree (in white) is superimposed onto a cloudogram derived from 20,000 post-burn-in trees (see [8]). Support was 100% for the two internal nodes closest to the root and very low within the capuchino radiation. Additional phylogenetic analysis indicates that the seven species (excluding the two outgroups, *S. bouvreuil* and *S. minuta*) radiated within the last 1.2 million years, and they are equally divergent from one another [6]. **(B)** The *Setophaga* (previously *Dendroica*) North American wood warblers. A phylogeny estimated for *Setophaga* species from multiple nuclear genes [34]. Support for internal nodes is color-coded: warm – high, cold – low. Maps indicate geographic ranges for taxa in two clades ((i) and (ii)). Note that ranges exhibit different degrees of overlap and three of the taxa in clade (ii) are island endemics, which is likely to reduce gene flow with other taxa.

Demography inference

Our study of the two avian groups will be based on inference of the demographic history of species in these groups and examination of different hypotheses regarding their divergence. Our aim is to generate a single unified analysis for each group that will generate comprehensive joint inference for all speciation events in that group. However, because the two data sets are expected to be very large ($\sim$10 populations and 3-5 individuals per species), we will explore these data sets initially by subsetting the data. In some analyses we will subsample the inidividuals per population and cross validate our inferernce across runs. We expect the results to be robust to the chnages in the identity of the individuals analyzed. In other analyses we will subset the species and conduct separate runs on overlapping subsets. We implemented this strategy in the tripplet validation we conducted in the preliminary analysis and showed it to be useful to validate our main findings (see **Preliminary results**). These subsampling strategies will be mostly used for initial exploration and validation, and we will strive to base our main conclusion on analyses that cover as much of the data as possible.

Role of gene flow

One of the primary objectives of our study is to elucidate the role of gene flow in speciation in these two groups. To achieve this we will use relative Bayes factors to compare various plausible hypotheses regarding scenarios of divergence and gene flow in the two avian groups. Firstly, we are interested in determining and characterizing the admixture events in these groups. This can be particularly challenging when the topology of the population phylogeny is not fully resolved, as is the case with these groups (see Figure 2 and **Preliminary results**). We will address this challenge by using relative Bayes factors to examine plausible topologies for the population phylogeny and various collections of migration bands. We will use insights obtained from our simulation study to guide our choice of reference model for these tests. Another central issue we would like to address using the model comparison framework is determining whether a certain admixture event is more compatible with continuous gene flow or secondary contact. Currently, G-PhoCS models gene flow continuously, but the coalescent model can easily be modified to assume a model of secondary contact after divergence. Relative Bayes factors can thus be used to help us distinguish between these two alternative hypotheses.

Natural selection

The main objective of our empirical study is to characterize neutral patterns of variation along the genome, as we expect nearly all sequence analyzed to be neutrally evolving. However, these neutral patterns provide some information about the influence of natural selection on species divergence. For instance, if closely related species co-exist in allopatry but exhibit low levels of gene flow, this could be used to indicate a reproductive barrier forming between the two species. To further explore the influence of natural selection on gene flow, we will attempt to assess whether gene flow is allowed in all loci at the same rate (the default model) or whether gene flow is prohibited in a certain unknown subset of the loci. Such comparisons will play an important role in determining the role of selection on patterns of divergence along the genome. However, we note that this is a fairly subtle comparison, since the two models could be very similar in cases with low rates of gene flow. We will thus use simulated data to determine how high the migration rate needs to be and how much data is required to tell these two models apart. We note that

3.3 Research design and methods – tool development

We will produce software tools for wide community use based on the methods developed, tested, and applied in this study. We will focus on user-friendly modules for processing the output traces of the MCMC. These modules, implemented in Java, will guide users in setting up the analysis, examining and selecting adequate reference models, and ranking of plausible hypotheses. A dedicated software developer will be employed to develop and maintain these tools. The programmer will work in close contact with our collaborators from the Lovette lab to ensure that the implemented features suit the needs of evolutionary biologists.

3.4 Preliminary results

Preliminary whole-genome sequence data was obtained for *Sporophila* by the Lovette laboratory, and we published our initial findings earlier this year [7]. The analysis focused on designing an adequate pipeline for demography inference using ddRADtag data, and comparing the coalescent-based inference of G-PhoCS to population structure and admixture analysis, which is the common standard approach used in such studies.

Sequence data. Sequencing libraries for ddRADtag were prepared using the protocol described by [40], and sequenced on an Illumina HighSeq 2500 machine to produce 150 basepair (bp) reads for 106 *Sporophila* individuals from nine species (see Fig. 2A, excluding the distant outgroup *Sporophila minuta*). The reads were trimmed to 125 bp and subsequently filtered for quality and adapter contamination. Using the stacks pipeline [10], roughly 320 million high quality filtered reads were clustered into approximately 86,000 genomic loci of length 125 bp, resulting in total coverage of roughly 1% of the *Sporophila* genome. Due to the stochastic nature of the restriction enzyme used in the ddRADtag protocol, for each locus we have sequence information only from a subset of individuals. We thus excluded three of the nine species (*S. cinnamomea*, *S. hypochroma* , and *S. nigrorufa*) from the preliminary demography analysis due to relatively sparse data, and then sub-selected a set of 3,763 loci where we had sequence information for at least one individual for each of the remaining six species.

Demography inference. We conducted demography inference by applying G-PhoCS to the sequence alignments for eighteen individuals (three per species) at the 3,763 filtered loci. We assumed a population phylogeny in which *S. bouvreuil* was an outgroup and the remaining five taxa were children of a five-way multifurcating node, consistent with the lack of phylogenetic resolution within that clade (see Fig. 2A). The phylogeny was augmented with 32 directional migration bands representing all possible branch pairs, including ancient migration between *S. bouvreuil* and the population ancestral to all other five. We ran G-PhoCS v.1.2.2 using the standard settings for the MCMC described previously in [20, 19], with 100,000 burn-in iterations and 200,000 additional sampling iterations. The resulting posterior estimates are summarized in Figure 3A. We infer that the outgtoup diverged roughly 1.85 million generations ago (ga) (1.77 – 1.94 million ga; 95% Bayesian credible interval; see Fig. 3A), and the remaining five species diverged from each other much after that, at 44 thousand ga (10 – 25 thousand ga). This provides strong support to previous findings of higher divergence for *S. bouvreuil* [8]. **The most striking finding is that the divergence of the five species is associated with a ∼10-fold increase in population size from roughly 1.45 million to over 14.4 million after their divergence from the outgtoup, followed by a sharp reduction in size after their subsequent divergence from each other**. Additionally, we infer relatively high levels of gene flow between *S. bouvreuil* and the other five species, including ancestral gene flow at a rate of roughly 25% between *S. bouvreuil* and the ancestral population. We did not detect significant levels of gene flow between the other five species, likely due to their very recent divergence and the large ancestral population size.

Validation through subset analysis. For validation, we ran G-PhoCS an additional ten times using species triplets, with *S. bouvreuil* as an outgroup (Fig. 3B), and compared estimates of shared parameters across these ten runs with our main run. Most paramaeters showed high levels of consistency, as shown for the outgroup divergence time in Figure 3C (τ_1). We take this to imply that our estimates are robust and not influenced by artifacts that may influence individuals of particular species. The parameter that showed the highest variance in estimate values was the divergence time between the ingroup species (τ_2). Importantly, this broad range of estimates still supports our hypothesis of recent radiation and the fact that *S. bouvreuil* is a clear outgroup to the other species. We take this finding to suggest a complex speciation process involving these five *Sporophila* species.

Population structure analysis. To compare our findings against those obtained by other commonly used inference tools, we conducted a population structure analysis using a popular software tool called *STRUCTURE* (version 2.3.4; [43]). We implemented an admixture ancestry model with K=1 through 10 clusters, conducting ten iterations per K value, and found that K=6 had the highest support. The cluster coefficients (Fig. 3C) show a large proportion of about 90% of shared ancestry to all individuals. This is consistent with the very large ancestral population size inferred by G-PhoCS, and the high rates of gene flow from the outgroup. Other than that, there appears to be a cluster assigned to the outgroup species (*S. bouvreuil*; red), but the other four clusters do not clearly correspond to the any of the five species. Thus, *STRUCTURE* appears to have weaker power than G-PhoCS in differentiating between these recently diverged species.

Conclusion. Our preliminary analysis demonstrates that G-PhoCS can conduct robust demography inference from ddRADtag data. In future analysis, we will relax our rigorous filtering scheme, which we expect to considerably increase the amount of information and further reduce in size of the credible intervals. The large ancestral population size and prevalent gene flow confirm our initial hypothesis regarding complex speciation in *Sporophila*. Our analysis also demonstrates the advantage of using a model-based method, such as G-PhoCS, compared to other popular methods in population genetics, such as STRUCTURE. We expect that the Bayesian hypothesis testing framework outlined in Section 3.1 will help us uncover the evolutionary history of these species in much greater detail and accuracy.

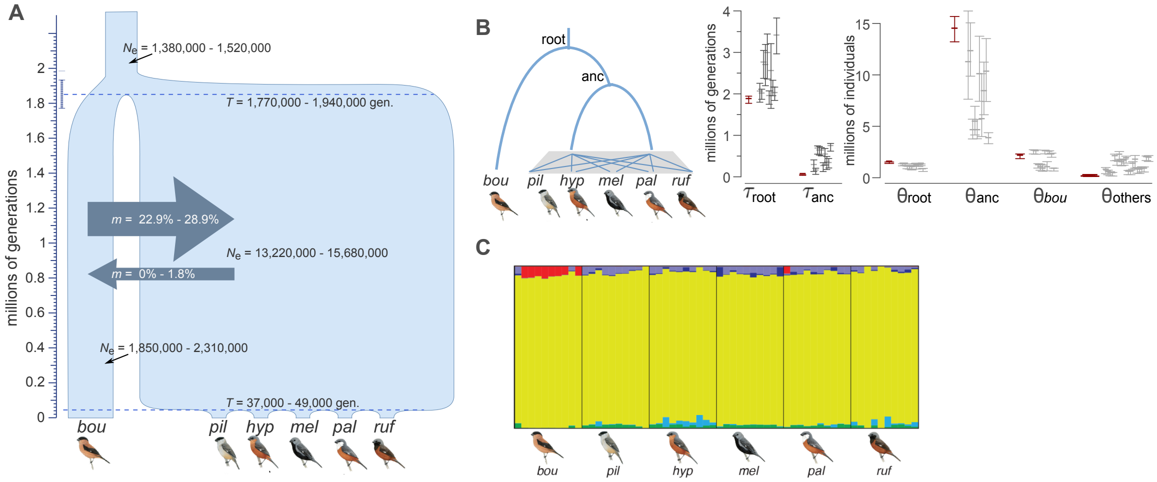

Figure 3. Preliminary demography inference for *Sporophila*. (A) Demography inference was conducted by applying G-PhoCS to the multiple sequence alignments of 18 individuals from six *Sporophila* species at 3,763 genomic loci of length 125 bp. The figure specifies posterior estimates of effective population sizes, divergence times, and total rates of migration (the product of the per-generation migration rate and the number of generations gene flow is assumed to take place). Ranges correspond to 95% Bayesian credible intervals and the two dashed horizontal bars represent the posterior mean of the two divergence times. Parameter values are converted from mutation scale to generations (τ) and individuals (θ) by assuming an average mutation rate of 10^{-9} mutations per bp per generation [31]. Note the dramatic increase in population size associated with the divergence of the five species from the outgroup, followed by a dramatic reduction after their divergence from each other. Significant rates were only inferred for flow between the outgroup (*S. bouvreuil*) and the other five species, especially with the ancestor of all five species. **(B)** A series of small-scale G-PhoCS analyses using pairs of sister taxa and *S. bouvreuil* as an outgroup. Estimates are shown for the common model parameters across the 10=$\binom{5}{2}$ separate runs (each run is represented twice in the plots), alongside the estimate from our main analysis (red bars; see panel A). Estimates are fairly consistent across runs, with fluctuations in estimates of τ_{anc} indicating a complex speciation process for the five species that took place throughout a time period of roughly 50,000 generations. **(C)** Bayesian admixture inference using state-of-the-art population genetic software tools. We used *STRUCTURE* version 2.3.4 [42] to assign individuals to genetic clusters. We implemented the admixture ancestry model and correlated allele frequencies, exploring values of K = 1 through 10 (conducting ten iterations per K value). Each run consisted of 300,000 generations following a burn-in of 200,000. The most likely K value was determined following methods described in [17] and implemented in *STRUCTURE* Harvester version 0.6.94 [16]. Different iterations from the optimal K value were combined in Clumpp version 1.1.2 [27] and displayed graphically using Distruct version 1.1 [47]. The dominant admixture component in all species indicates shared ancestry consistent with the large ancestral population inferred using G-PhoCS. The structure analysis appears to have weak power distinguishing the outgroup from the other five species (small red component), and practically no power to distinguish between the five ingroup species. Figures taken from [7] published this year.

3.5 Research facilities

I intend to conduct the proposed research at the Efi Arazi School of Computer Science at the Interdisciplinary Center (IDC) in Herzliya. The school provides a fully equipped computer lab, where programmers and students will develop software. It also provides workspace for students and postdocs to conduct theoretical research. I plan to supplement the existing equipment in the lab with a Linux server where data will be stored and large compute jobs will be run (see detailed budget proposal). The study will also take advantage of the strong collaboration established with the Lovette laboratory at the Cornell Lab of Ornithology. The Lovette lab is equipped with state-of-the-art sequencing facilities, and is committed to generating sequence data to promote this research project.

3.6 Expected outcome and pitfalls

Methods. One of the major objectives of this research is to address the methodological challenges in examining and scoring different evolutionary hypotheses. By combining the new theoretical ideas outlined in Section 3.1 with rigorous testing on simulated data, we expect to produce **effective methods and detailed recipes for investigating complex scenarios of gene flow**. We see two main sources of complication. The first source of complication has to do with selecting an appropriate reference model to use in model comparison. As we note in Section 3.1, the default panmictic model might not be sufficiently similar to the models being compared, leading to large variance in the estimation of the relative Bayes factors. That, in turn, could bias our results. We thus intend to use extensive simulations to examine effective strategies for selecting the reference model. Furthermore, if estimation bias remains high, we will use variance reduction techniques, such as thermodynamic integration [37], to improve accuracy with an additional computational cost. The second source of potential complication is that the genealogy summaries that we intend to use to suggest alternative models (see Section 3.1) will be too noisy and thus only weakly informative about lack of model fit. To address this we intend to start with simple cases of two-population tests that have been proven to work in previous studies [15], and gradually increase the level of complexity by adding more individuals and populations.

Data analysis. Through analysis of sequence data from *Sporophila* and *Setophaga*, we expect to reach **new general insights on the demographic processes involved in species radiation**. Our preliminary analysis suggests that the proposed computational approach is adequate, sufficiently powerful, and serves as a considerable improvement to current popular approaches. By applying the methodological extensions to this data, and future data generated by our collaborators, we expect to create a powerful demonstration for how such studies should be carried out. Our goal is to reliably reconstruct a detailed account for the demographic processes associated with species radiation in these two groups. This includes best fit hypotheses for scenarios of gene flow. One possible complication is the sequencing method used. We are currently relying on ddRADtag data, which has several key weaknesses: (1) we do not have information about the location of the analyzed loci, and (2) the sampling of loci is not entirely random and is biased toward less variant loci due to the presence of a conserved restriction site. We thus plan to test this approach by generating complete genome sequence data for a small number of individuals in each of the two group, and comparing our ddRADtag-based inference with that based on complete genome sequences. Eventually, we intend to combine the two types of data to take advantage of the relative strengths of each of them.

Public software. We will implement all computational methods we develop for this study as open source software. In addition to providing fully documented source code, we will develop a complete user-friendly interface that will facilitate the use of these heavy computational tools by a community of evolutionary biologists. We will supplement this with a web server that will enable researchers to run small scale analyses on our Linux server (see http://compgen.cshl.edu/INSIGHT for an example of such a webserver).

References

1. Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

2. Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics. 162:2025–2035.

3. Beerli P, Felsenstein J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics. 152:763–773.

4. Benites P, Campagna L, Tubaro PL. 2014. Song-based species discrimination in a rapid Neotropical radiation of grassland seedeaters. Journal of Avian Biology. .

5. Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol. 25:1979–1994.

6. Campagna L, Benites P, Lougheed SC, Lijtmaer DA, Di Giacomo AS, Eaton MD, Tubaro PL. 2012. Rapid phenotypic evolution during incipient speciation in a continental avian radiation. Proc. Biol. Sci. 279:1847–1856.

7. Campagna L, Gronau I, Silveira LF, Siepel A, Lovette IJ. 2015. Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. Mol. Ecol. 24:4238–4251.

8. Campagna L, Silveira LF, Tubaro PL, 2, Lougheed SC. 2013. Identifying the Sister Species to the Rapid Capuchino Seedeater Radiation (Passeriformes: Sporophila). The Auk. 130:645–655.

9. Carbone L, Harris RA, Gnerre S, et al. (93 co-authors). 2014. Gibbon genome and the fast karyotype evolution of small apes. Nature. 513:195–201.

10. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. 22:3124–3140.

11. Coyne JA. 1992. Genetics and speciation. Nature. 355:511–515.

12. Coyne JA, Orr AH. 1989. Patterns of speciation in Drosophila. Evolution. 43:362–381.

13. Coyne JA, Orr AH. 1990. Hybrid zones: Windows on evolutionary process. Oxford Surveys in Evolutionary Biology. 7:69–128.

14. Csillery K, Blum MG, Gaggiotti OE, Francois O. 2010. Approximate Bayesian Computation (ABC) in practice. Trends Ecol. Evol. (Amst.). 25:410–418.

15. Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.

16. Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources. 4:359–361.

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

18. Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.

19. Freedman AH, Gronau I, Schweizer RM, et al. (30 co-authors). 2014. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. 10:e1004016.

20. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43:1031–1034.

21. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695.

22. Hastings WK. 1970. Monte carlo sampling methods using markov chains and their applications. Biometrika. 57:97–109.

23. Hey J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27:905–920.

24. Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics. 167:747–760.

25. Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. U.S.A. 104:2785–2790.

26. Hudson RR. 1991. Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J, editors, Oxford Surveys in Evolutionary Biology, volume 7, pp. 1–44.

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

28. Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro H, editor, Mammalian Protein Metabolism, New York: Academic Press, pp. 21–132.

29. Kimura M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

30. Kuhner MK, Yamato J, Felsenstein J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics. 140:1421–1430.

31. Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. Proc Natl Acad Sci USA. 99:803–808.

32. Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.

33. Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B. 2013. Efficient moment-based inference of admixture parameters and sources of gene flow. Mol. Biol. Evol. 30:1788–1802.

34. Lovette IJ, Perez-Eman JL, Sullivan JP, et al. (12 co-authors). 2010. A comprehensive multilocus phylogeny for the wood-warblers and a revised classification of the Parulidae (Aves). Mol. Phylogenet. Evol. 57:753–770.

35. McManus KF, Kelley JL, Song S, et al. (12 co-authors). 2015. Inference of gorilla demographic and selective history from whole-genome sequence data. Mol. Biol. Evol. 32:600–612.

36. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller A, Teller E. 1953. Equation of state calculations by fast computing machines. J Chem Phys. 21:1087–1092.

37. Newton MA, Raftery AE. 1994. Approximate bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society. Series B (Methodological). pp. 3–48.

38. Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics. 158:885–896.

39. Nosil P, Feder JL. 2012. Widespread yet heterogeneous genomic divergence. Mol. Ecol. 21:2829–2832.

40. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE. 7:e37135.

41. Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8:e1002967.

42. Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. 20:R208–215.

43. Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

44. Rabosky DL, Lovette IJ. 1999. Explosive speciation in the New World Dendroica warblers. Proc. R. Soc. Lond. B. 266:1629–1636.

45. Rabosky DL, Lovette IJ. 2008. Density-dependent diversification in North American wood warblers. Proc. Biol. Sci. 275:2363–2371.

46. Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.

47. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. Science. 298:2381–2385.

48. Shafer AB, Gattepaille LM, Stewart RE, Wolf JB. 2015. Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: in silico evaluation of power, biases and proof of concept in Atlantic walrus. Mol. Ecol. 24:328–345.

49. Sousa V, Hey J. 2013. Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. 14:404–414.

50. Toews DP, Mandic M, Richards JG, Irwin DE. 2014. Migration, mitochondria, and the yellow-rumped warbler. Evolution. 68:241–255.