3 Detailed Description of the Proposed Research

Achieving the objectives listed above requires developing new theory and methods alongside extensive data analysis. A detailed outline of our plans in these two parallel tracks is described in Sections 3.1 and 3.2 below. We follow by summarizing our preliminary results (Section 3.3) and providing additional details about our research plan (Sections 3.4 and 3.5).

3.1 Research design and methods – theory and computational methods

The Genealogy Sampler

We start by providing a very brief description of the probabilistic model underlying the genealogy sampler of G-PhoCS (for more details, see [16]). The demographic model is based on a population phylogeny augmented by horizontal edges called *migration bands* (Fig. 1A). We denote the augmented population phylogeny by T and the demographic model by $M = (T, \Theta)$, where Θ denotes the set of free parameters, which consist of divergence times (τ) associated with internal nodes in T, effective population sizes (θ) associated with its branches, and migration rates (m) associated with migration bands. G-PhoCS infers the free parameters by fixing T and examining patterns of sequence variation in an alignment, $\mathcal{X}$, of multiple genomes at L short loci assumed to be genetically unlinked and neutrally evolving (Fig. 1B). The likelihood function, $\mathcal{L}$, involves associating each locus $l = 1..L$ with a local genealogy (or gene tree), G_l, and is given by a product across loci of the genealogy prior, $P(G_l|M)$, defined using coalescent theory [21, 1, 32], and the data likelihood, $P(X_l|G_l)$, defined using a DNA substitution model (e.g. [23, 25, 14]).

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

G-PhoCS assumes a weak prior distribution for the model parameters, $P(\Theta)$, and jointly samples values of the parameters and instances of the local genealogies according to an approximate posterior distribution, $P(\Theta, \{G_l\}|\mathcal{X}, T) = P(\Theta)\ \mathcal{L}(T, \Theta, \{G_l\})/P(\mathcal{X}|T)$ (Fig. 1C). This Bayesian inference scheme is implemented using a Markov Chain Monte Carlo (MCMC) sampling algorithm [30, 18].

Genealogy summaries

The current version of the sampler provides robust estimates of demographic model parameters conditioning on a given model, T. However, it does not provide adequate means to explore different models and assess the fit of the model to the data. To enable refinement of the assumed model, we suggest using summaries of the sampled genealogies that indicate lack of fit. This approach has been recently suggested as simple means of examining gene flow and population phylogeny topologies [11, 35, 28]. However, explicit utilization of the sampled genealogies is the most direct way to achieve this. By recording features of the sampled genealogies that deviate from the assumed model, we will obtain information on model assumptions that are wrong. 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. The D-statistic [11] does this for the simple case of four lineages and one possible gene flow event. Using the more general setup of the genealogy sampler will allow us to extend this test to a larger number of lineages and more complex scenarios of gene flow.

Bayesian model comparison

Evolutionary inference requires being able to distinguish between different hypotheses regarding the evolutionary history of the analyzed samples and indicating which of them has more support in the data. In our framework, these hypotheses are translated into assumptions regarding the structure of the population phylogeny, scenarios of gene flow, and ranges of demographic parameters in Θ. Thus we associate hypotheses with different assumptions on $M = (T, \Theta)$ and would like to score a given model, M, proportionally to its fit to the data, $P(X|M)$. The Bayesian approach provides a natural way to achieve this through the concept of *Bayes factors*. Let us denote by $\mathcal{G} = (\Theta, \{G_l\})$ an assignment to the demographic parameters and all local genealogies, as sampled by G-PhoCS assuming model M. The Bayes factor of model M, $P(\mathcal{X}|M)$, can be computed as follows:

$$\begin{aligned}\frac{1}{P(\mathcal{X}|M)} &= \int \frac{P(\mathcal{G}|M)}{P(\mathcal{X}|M)} d\mathcal{G} = \int \frac{P(\mathcal{G}|M)P(\mathcal{G},\mathcal{X}|M)}{P(\mathcal{G},\mathcal{X}|M)P(\mathcal{X}|M)} d\mathcal{G} = \int \frac{P(\mathcal{G}|\mathcal{X},M)}{P(\mathcal{X}|\mathcal{G},M)} d\mathcal{G} \\ &= \mathbb{E}_{\mathcal{G}|\mathcal{X},M}\left(\frac{1}{P(\mathcal{X}|\mathcal{G},M)}\right), \qquad (2)\end{aligned}$$

where $\mathbb{E}_{\mathcal{G}|\mathcal{X},M}$ represents the expectation under the posterior distribution of $\mathcal{G}$ given $\mathcal{X}$ and M. Thus, the Bayes factor, $P(\mathcal{X}, M)$, can be approximated by taking the *harmonic mean* of $P(\mathcal{X}|\mathcal{G}, M) = \prod_l P(X_l|G_l)$ across K samples of the MCMC. The harmonic mean provides a simple and general means of approximation for the Bayes factor, but it is notoriously difficult to apply to real data, because it is an unstable estimator [31]. Various computational techniques have been proposed to improve its statistical efficiency [27], but those require an additional computational load, which is not feasible in our case. We are proposing to bypass this problem by estimating the ratio between the Bayes factor of M and another model M_0, based on the following observation:

$$\begin{aligned}\frac{P(\mathcal{X}|M_0)}{P(\mathcal{X}|M)} &= \int \frac{P(\mathcal{G},\mathcal{X}|M_0)}{P(\mathcal{X}|M)} d\mathcal{G} = \int \frac{P(\mathcal{G},\mathcal{X}|M_0)P(\mathcal{G},\mathcal{X}|M)}{P(\mathcal{G},\mathcal{X}|M)P(\mathcal{X}|M)} d\mathcal{G} = \int \frac{P(\mathcal{G},\mathcal{X}|M_0)}{P(\mathcal{G},\mathcal{X}|M)} P(\mathcal{G}|\mathcal{X},M) d\mathcal{G} \\ &= \int \frac{P(\mathcal{X}|\mathcal{G})P(\mathcal{G}|M_0)}{P(\mathcal{X}|\mathcal{G})P(\mathcal{G}|M)} P(\mathcal{G}|\mathcal{X},M) d\mathcal{G} = \int \frac{P(\mathcal{G}|M_0)}{P(\mathcal{G}|M)} P(\mathcal{G}|\mathcal{X},M) d\mathcal{G} \\ &= \mathbb{E}_{\mathcal{G}|\mathcal{X},M}\left(\frac{P(\mathcal{G}|M_0)}{P(\mathcal{G}|M)}\right). \qquad (3)\end{aligned}$$

This implies the following scheme for estimating the relative Bayes factor for a given collection of models M_1,..., M_N: (1) choose a model M_0 for which $P(\mathcal{G}|M_i) > 0 \rightarrow P(\mathcal{G}|M_0) > 0$ for all $i = 1 \ldots N$; (2) for each i, run the genealogy sampler under model M_i to generate K samples $\{\mathcal{G}_k^i\}$ under the approximate porterior $P(\mathcal{G}|\mathcal{X}, M_i)$; (3) estimate the relative Bayes factor $\frac{P(\mathcal{X}|M_0)}{P(\mathcal{X}|M_i)} \approx$

$BF(M_0 : M_i) = \frac{1}{K}\sum_k \frac{P(\mathcal{G}_k^i|M_0)}{P(\mathcal{G}_k^i|M_i)}$. This is schematically demonstrated in Figure 2. Despite the fact that this approach does not provide an absolute estimate of the Bayes factor, $P(\mathcal{X}|M_i)$, for any of the N models, it provides estimates for all ratios $P(\mathcal{X}|M_i)/P(\mathcal{X}|M_j) \approx BF(M_0 : M_j)/BF(M_0 : M_i)$. Additionally, this approach is expected to be more statistically efficient than the harmonic mean estimator because the values it is averaging over, $\frac{P(\mathcal{G}_k^i|M_0)}{P(\mathcal{G}_k^i|M_i)}$, have a much lower variance than their counterparts, $\frac{1}{P(\mathcal{X}|\mathcal{G}_k^i, M_i)}$.

Testing and validation

Based on the approach outlined in the previous section, we will implement methods that use the genealogy sampler to propose, test, and compare different hypotheses regarding the evolutionary history of the analyzed populations. We will use extensive coalescent simulations to test these methods and their power to distinguish between different models and detect lack of fit. These methods will be integrated into G-PhoCS, and we will use them to re-examine publicly available data [16, 15, 5, 24]. We will make use of our familiarity with these data to draw conclusions both on the developed methods and the species being examined. We will also continue to extend and improve the current sampler and improve its computational efficiency by optimizing the source code and introducing parallelization.

3.2 Research design and methods – avian 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 *Satophaga* 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 [38, 39, 29, 4]. 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., [2, 44])

- 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. 3).

- 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 Section 3.3).