# Riding the speciation continuum: discordance between genomic divergence and phenotypic variation in a rapidly evolving avian genus (*Motacilla*)

Rebecca B. Harris

A dissertation

submitted in partial fulfillment of the

requirements for the degree of:

Doctor of Philosophy

University of Washington

June 2017

Reading Committee:

Adam D. Leaché, Chair

John Klicka

Sharlene Santana

Program Authorized to Offer Degree:

Biology

University of Washington

## Abstract

Riding the speciation continuum: discordance between genomic divergence and phenotypic variation in a rapidly evolving avian genus (*Motacilla*)

Rebecca B. Harris

Chair of the Supervisory Commitee:

Adam D. Leaché

Department of Biology

Investigating heterogeneity in genomic divergence in species with geographically variable phenotypes may reveal the evolutionary processes responsible for speciation. Generally, genotypes and phenotypes are expected to be spatially congruent; however, in widespread species complexes with few barriers to dispersal, multiple contact zones, and limited reproductive isolation, discordance between phenotypes and phylogeographic groups is more probable. Wagtails (Aves: *Motacilla*) are a genus of birds with striking plumage pattern variation across Eurasia. Up to 13 subspecies are recognized within a single species, yet previous studies using mitochondrial DNA have supported phylogeographic groups that are inconsistent with subspecies plumage characteristics. In Chapter 1, we investigate the link between phenotypes and genotype by comparing populations thought to be at different stages along the speciation continuum. We take a phylogeographic approach by estimating population structure, testing for isolation by distance, conducting demographic modeling, and estimating the first time-calibrated species tree for the genus. Chapter 1 provides strong evidence for

species-level patterns of differentiation in wagtails, however population-level differentiation is less pronounced. We find evidence that three of four widespread Eurasian species exhibit an east-west divide that contradicts both subspecies taxonomy and phenotypic variation. Both the geographic location of this divide and time estimates from demographic models are overlapping in two sympatric species, indicating that coincident Pleistocene events shaped their histories.

The study of non-model organisms is often hampered by lack of a reference genome. A physical linkage map improves overall data quality, decreases the false-positive rate of genome scans, and improves the statistical power of methods to detect selection. In Chapter 2, we provide a 1.1 Gb reference genome for *Motacilla flava* based on the *de novo* assembly of whole-genome shotgun and long-insert mate-pair libraries. Annotation was conducted using the flycatcher cDNA library. Use of a reference genome has markedly improved assembly of wagtail ddRAD sequencing data (Chapter 1) and low-coverage whole genome shotgun data (Chapter 3). In Chapter 3, we use whole-genome sequencing to confirm the population genetic structure and demographic estimates found in Chapter 1, as well as to provide a platform for future studies on selection.

**ACKNOWLEDGEMENTS**

I thank the museums and those who sent me samples: U. Johansson, Naturhistoriske Riksmusect; B. Schmidt, USNM; I. Nishiumi, NMSM; B. Marks, FMNH; Mark Robbins, KU; Jack Withrow, UAM; P. Sweet, AMNH; K. Zyskowski, YPM; M. Westberg, Bell Museum; G. Voelker, TCWC; S. Birks, UWBM. I thank M. Melo for sharing tissues and D. Shizuka facilitating contacts in Japan. For helpful discussion and comments on Chapter 1, I thank S. Rohwer, J. Klicka, N. Sly, S. Billerman, E. Linck, and C. Battey. For assistance with sequencing and analysis, I thank R. Gutenkunst, N. Bouzid, D. Petkova, and J. Grummer. I thank D. Scofield for his guidance on Chapter 2.

I was supported by the NSF-GRFP. Funding came from the British Ornithologists' Union, AMNH Chapman Grant, Sigma Xi GIAR, NSF DDIG DEB-1501131, UW Sargent Award, WRF Hall Fellowship, and the Burke Museum of Natural History and Culture.

I used the UC Berkeley Vincent J. Coates Genomics Sequencing Laboratory, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303, and the advanced computational, storage, and networking infrastructure provided by the University of Washington Hyak supercomputer system. I acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure.

**Chapter 1:** Riding the speciation continuum: discordance between genomic divergence and phenotypic variation in a rapidly evolving avian genus (*Motacilla*)

Rebecca B. Harris[1,2], Per Alström[3,4,5], Anders Ödeen[3], and Adam D. Leaché[1,2]

[1] Department of Biology, University of Washington, Seattle, WA 98195

[2] Burke Museum of Natural History and Culture, Seattle, WA 98195

[3] Department of Animal Ecology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[4] Swedish Species Information Centre, Swedish University of Agricultural Sciences, Box 7007, Uppsala SE-750 07, Sweden

[5] Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

*We dedicate this work to Anders Ödeen, a dear friend, colleague, and pioneer in wagtail genetics, who passed away during the preparation of this manuscript.*

## Introduction

Species complexes are often characterized by high frequencies of hybridization and poorly developed isolation barriers, despite being structured geographically [1]. This can lead to incongruence between genotypes and phenotypes [2]. Lack of overall genetic differentiation in taxa with distinct phenotypic differences is likely due to either (1) recent divergence, with strong selection on phenotype, or (2) large-scale introgression, except on pre-existing adaptive genetic differences. Non-adaptive demographic processes will have a universal and random effect on the genome [3], whereas selection may cause differentiation at few adaptive loci [4]. Therefore, early in the speciation process, patterns of genetic divergence may be inconsistent with the history of population divergence [5].

In birds, plumage differences among closely related species are often believed to be the result of sexual selection and to play an important role in reproductive isolation [1]. Plumage differences can evolve rapidly [6–8] and, when populations are geographically structured, may result from spatial variation in selection regimes [1]. Recent studies have demonstrated that a small number of genes can cause dramatic plumage differences despite limited genetic differentiation throughout the remainder of the genome [2,9–11].

Widespread species complexes are useful systems for investigating heterogeneity in genomic divergence and geographic variation in phenotypes because they offer comparisons between populations at different stages of the speciation continuum. Stages in the speciation process, from barely differentiated allopatric/parapatric populations, via subspecies/species with distinct plumages that meet in variously wide hybrid zones, to fully reproductively isolated species [12]. Moreover, plumage differences are thought to have evolved rapidly and in conflict with phylogeographic structure [13–17].

One bird system that is particularly well suited for such studies is the passerine genus *Motacilla*. The genus *Motacilla* consists of 12 species distributed throughout the Old World [12,18]

that have earned the common name wagtail due to their propensity to pump their long tails up and down. Striking examples of spatial variation in plumage traits are found in the *M. alba* (9 subspecies) and *M. flava* (13 subspecies) complexes (cf. Fig. 1) and many of these have been treated as separate species (reviewed in Alström & Mild 2003). Both species complexes exhibit high levels of geographic variation, especially in male breeding plumage; subspecies are defined by head plumage color (ranging from white to green and yellow to black) and pattern in the *M. flava* complex and by head, upper parts, and wing-covert color/pattern in *M. alba*. In contrast, *M. cinerea* (three subtle subspecies, two of which are insular) and *M. citreola* (two subspecies differing in upper part color) lack this extreme plumage variation. These four species are widely distributed and sympatric across the Palearctic during the breeding season and are long-distance migrants at least in the northern parts of their respective ranges. Three species are monotypic and have rather restricted allopatric distributions in the Indian subcontinent (*M. maderaspatensis*), Cambodia (*M. samveasnae*), and Japan (*M. grandis*), whereas the three Afrotropical (*M. aguimp*, *M. capensis*, *M. clara*) and single Malagasy species (*M. flaviventris*) exhibit slight or no geographical plumage variation and are resident [12,18]. Previous phylogenetic and phylogeographic studies of *Motacilla* report mitochondrial relationships incongruent with both taxonomy [13,15–17] and nuclear relationships [14,19,20], with suggestions that mitochondrial DNA (mtDNA) poorly reflects the true phylogeny [12,14,19,20]. Several of these studies focused on aspects of wagtails' plumage diversity, some proposing cases of remarkable parallel plumage evolution [12,19,20] and others implicating the role of selection in rapid plumage evolution [13]. A more recent phylogenetic exploration of the genus found the São Tomé endemic *Amaurocichla bocagii* nested within *Motacilla*, and proposed its inclusion within the genus [21].

Much of wagtail divergence is thought to have occurred during the Pleistocene, either through dispersal [15] and/or mediated by climatic events [17]. Climate changes caused by cyclical glaciation and aridification during the late Pliocene through the Pleistocene had drastic effects on many modern day species distributions [22–24]. The ancestors of modern populations would have had to seek refuge in isolated pockets of suitable habitat, potentially resulting in population differentiation due to genetic

drift and natural selection in isolation. Three major refugia are thought to have periodically existed in the eastern, western, and southern edges of Eurasia since the late Miocene [23].

While the extent to which climate oscillations affected distributions is species-specific [25], both *M. alba* and *M. flava* display geographic variation in plumage types (i.e., subspecies) beyond the number of plausible refugia [23]. The "excess" of *M. alba* and *M. flava* subspecies in relation to refugia could be due to waves of isolation corresponding with past climate oscillations, which were punctuated by periods of sympatry that reinforced sexual or social character displacement over the course of multiple cycles [26]. Individual wagtail species may have responded to these changes differently, or may have been impacted by the same events [27]. Alternatively, differentiation could have occurred after dispersal out of glacial refugia following the last glacial maximum, and due to rapid local adaptation or local sexual selection across the wide distribution of these birds.

In this study, we utilize genome-wide SNPs, nuclear introns, and mtDNA to analyze speciation patterns in *Motacilla*, with complete species-level sampling and comprehensive coverage of the three most diverse Palearctic species. We (1) estimate the first time-calibrated species tree for this group; (2) demonstrate conclusively that mtDNA alone is inappropriate for phylogenetic studies of *Motacilla*; (3) investigate the agreement between genotype and phenotype, along with the proposed parallel and rapid evolution of plumage in the three most variable wagtail species; and (4) model past demographic events to reveal the temporal dynamics of effective population size and investigate how past climatic events influenced *Motacilla*.

## Materials and methods

### *Sanger data*

To resolve species-level relationships within *Motacilla*, we utilize previously published and unpublished sequences from (1) three nuclear introns (*CHD1Z*, *ODC*, *Mb*) for 42 individuals across all 12 *Motacilla* species [14,19], and (2) two mitochondrial regions (*ND2, CR*) for 103 individuals

across all species, including all subspecies of *M. alba, M. flava,* and *M. citreola* [13–15,17]. *Dendroanthus indicus, Anthus pratensis*, and *Anthus trivialis* were used as outgroups [21].

*ddRADseq data*

*Sampling* - If wagtail divergence was recent or shaped by rapid ancestral radiations, then the timing between divergence events may have been too short for the emergence of phylogenetically informative mutations [28,29], potentially leading to the mito-nuclear discordance shown in previous studies. We therefore enhanced our inferential power by collecting thousands of genome-wide SNPs.

Throughout the manuscript, we follow the taxonomy of Alström and Mild (2003) and Alström *et al.* (2015) [12,21]. We obtained extensive geographic sampling and near-complete taxonomic coverage across Eurasia from samples at the Burke Museum of Natural History and Culture. We augmented these specimens with samples from other natural history museums to provide complete sampling for the genus. A total of 246 birds were sampled from 11 of the 12 recognized *Motacilla* species [12,21]. As the goal of our study is to elucidate processes responsible for population divergence, our sampling focused on the widespread migratory, Eurasian wagtail species (*M. alba, M. flava, M. citreola, M.cinerea*) with multiple described subspecies. We examined museum skins and assigned individuals to subspecies using morphological criteria outlined in Alström and Mild (2003). We were unable to include *M. maderaspatensis* in our SNP sampling due to a lack of available high-quality tissue samples. For rooting phylogenetic trees, we sampled three individuals from the monotypic sister genus (*Dendroanthus*) to serve as outgroups.

*Data collection* - We generated SNPs using the double-digest restriction site-associated DNA sequencing (ddRADseq) protocol following methodology described in Peterson *et al.* (2012) [30]. A total of five lanes were sequenced (single-end reads: four 50 bp and one 100 bp). We then constructed a reference genome for a single individual of *M. alba* to improve the accuracy of our ddRADseq locus assembly (see Chapter 2).

*Data filtering* - As missing data can affect downstream PCA results [31], phylogenetic inference [32,33], and other analyses as well [34], we further filtered datasets according to different levels of missing data (all possible combinations of 25%, 50%, and 75% missing loci and individuals). Together with our linkage filtering, we compiled 18 datasets per species group (two linkage treatments x three missing loci treatments x three missing individual treatments). We used all 18 datasets independently when conducting population structure analyses, Mantel tests, and RAxML phylogenetic trees. For all other analyses, we used a 50% threshold for both missing loci and individuals (see below).

*Phylogenetic analyses*

*Time-calibrated species tree and gene trees* - Nuclear DNA and mtDNA can support contradicting phylogenetic relationships and methods that ignore incomplete lineage sorting may fail to accurately estimate species tree relationships [35]. Because *Motacilla* lacks a species tree, we estimated the first time-calibrated species tree for this group using Sanger data (see above). We implemented molecular rate calibration in *BEAST v1.8.4 [36,37] using a published *Motacilla*-specific rate of 2.7% for *ND2* [17]. Recent simulation studies demonstrate that tree priors can have a large impact on divergence time dating when using datasets with mixed intra- and interspecies sampling [38]. Therefore, we conducted model selection on converged runs. As mtDNA and nuclear DNA may support different topologies for reasons other than incomplete lineage sorting, such as introgression and sex-biased dispersal, we conducted analyses on each data type independently. Using nuclear introns, we estimated a species tree in *BEAST. We also estimated separate nuclear and mtDNA gene trees in BEAST.

*Concatenated SNP tree* - To place subspecies in a wider phylogenetic context, we conducted ML phylogenetic analyses using concatenated ddRAD loci using RAxML v8.2 [39]. Due to their large population sizes and assumed recent divergences, wagtails contain a high level of heterozygosity with

few fixed SNPs: over 99% of all variable sites contain at least one individual with a heterozygous SNP. Most methods count heterozygous sites as missing data and researchers have typically excluded these sites. We implemented the method of Lischer *et al*. (2013) [40] to generate 500 random haplotype samples from sequences with multiple heterozygous sites. We then inferred ML phylogenies using RAxML for each of these datasets. To account for potential SNP ascertainment bias, we implemented the Felsenstein correction [41].

*SNP species tree* - To estimate a species tree, we implemented SNAPP [42]. SNAPP uses biallelic loci and requires at least one representative SNP from each species at each locus. Species assignments were based on population structure estimates (see below), and individuals with admixture were excluded to avoid model violations and branch length underestimation [43]. Because SNAPP is computationally intensive, convergence issues prevented us from using our full sampling scheme. We therefore ranked individuals by missing data and admixture proportions, and only included the top four from each population. All individual assignments were made with >95% posterior probability. Mutation rates (u, v) were both fixed at 1 and default parameters were used for the gamma prior (alpha 11.75, beta 109.73). Altering parameter values for the prior distributions on population size and tree length to better reflect wagtail population history caused convergence failure.

*Population genetic structure*

Grouping individuals based on phenotype may be misleading, especially given that current wagtail taxonomy does not reflect the mitochondrial or nuclear trees [12,20,44]. Objective, genetic-based methods are preferable for inferring the number of genetically distinct populations and the assignment of individuals to those populations. As published mtDNA-based studies suggest *M. flava* and *M. citreola* are polyphyletic [14,15,19,20,45], we initially ran all population structure analyses on these two species combined. A similar approach was taken for the "African" clade (*M. clara, M. capensis, M. flaviventris, M. bocagii*) and the "black-and-white wagtails" (*M. alba, M.*

*aguimp, M. samveasnae*, and *M. grandis*). Focal Eurasian wagtail species were further analyzed independently.

To *de novo* identify the optimal number of clusters in our data, we implemented two methods: the model free discriminant analyses of principal components (DAPC) in adegenet [46,47] and the model-based maximum-likelihood (ML) method, ADMIXTURE v1.3 [48]. One caveat of ADMIXTURE is that it assumes discrete ancestral or parental populations. When organisms exhibit continuous spatial population structure, recent studies tend to use PCA methods like adegenet [49]. However, adegenet has the undesirable behavior of assigning individuals to populations with unrealistically high probability. Therefore, to reduce the impact of their respective biases on downstream analyses, we employed these methods in concert to estimate *K* and individual assignment probabilities.

First, we ran ADMIXTURE with variable numbers of clusters *K*=1-10. We then plotted 10-fold cross-validation values terminated with default criteria, to choose the optimum value of *K*. Second, we ran the k-means clustering to assess groups using both AIC and BIC. We implemented DAPC to maximize differences between groups while minimizing variation within groups. To assess how many PCs to retain, we used cross-validation (*xvalDapc*) with 100 replicates and retained the number of PCs with the lowest mean squared error.

To assess genetic differentiation among phenotypes, we used DAPC to find the largest distance between subspecies defined *a priori*. For each species group, we ensured that a minimum of one individual per subspecies per locus was present. On average, this additional filtering step reduced our dataset by 40%. To determine the diagnosability of subspecies groups, we then compared our DAPC results to analyses with randomized subspecies definitions. Because rare alleles may be younger than common alleles, and may track more recent demographic or selective events [50], we explored their effect on population genetic inference by alternatively 1) pruning of all sites with minor allele frequency<10% or 2) building a matrix with only low-frequency alleles.

*Isolation by distance*

Populations in close proximity are expected to be more genetically similar than those located farther apart [51]. To explore whether our estimates are the result of the clustering of individuals with distinct allele frequencies or structure due to separation in space, we conducted Mantel tests on each species using the *mantel.randtest* function in ade4 [52] and ran these for 1 million permutations. To account for Earth's curvature, geographic distance was calculated using the Great Circle distance in sp [53].

Given that the ability of Mantel tests to detect isolation by distance (IBD) has been a recent area of debate [54,55], we also implemented the Estimated Effective Migration Surfaces (EEMS) method [56] to model the relationship between geography and genetics. EEMS allows visualization of variation in effective migration across each species' breeding distribution and the identification of corridors and barriers to gene flow by implementing a stepping-stone model over a dense grid. We used species' breeding ranges [57] and a dense grid of equally sized demes. To show that our results were independent of grid size, we ran each analysis using 250, 500, and 750 demes from three independent chains for 20 million MCMC iterations with a 10 million iteration burn-in. We first ran a series of short preliminary runs to choose parameter values that gave acceptance ratios between 20-30%. Graphs were constructed using rEEMSplots [56]. We visualized each run separately and checked convergence of MCMC runs (log posterior plots and Gewke diagnostic tests, Heidelberger and Welch test in coda [58]) before combining across runs and grid numbers to construct final consensus graphs.

*Demographic history*

To infer demographic history of each focal Eurasian species, we conducted ML parameter

estimation in *dadi* [59] using a joint allele frequency spectrum and our population structure estimates. To narrow the number of possible two-population (2D) models, we first optimized one-population models and compared support for instantaneous versus exponential population size change. Instantaneous change was preferred in all populations.

We next considered six alternative 2D models which offered simplified but reasonably realistic representations of plausible demographic processes. The simplest model is of allopatric divergence in which ancestral population ($N_a$) splits into two populations ($N_1$ and $N_2$) that evolve in isolation for $T_s$ generations. We then considered a scenario of divergence followed by instantaneous population change beginning $T_g$ generations ago and resulting in populations of size ($N_{1g}$ and $N_{2g}$). Finally, we incorporated asymmetric and symmetric gene flow into each of these base models. Since cycles of population expansions and contractions have been a common feature of many birds [24], we tried incorporating additional stages of population size change into our models. Unfortunately, we were limited by the size of our SNP matrix and we were unable to optimize these increasingly complex models.

All data were polarized using outgroup SNPs and an extra parameter was included in our 2D models to account for misidentified SNPs. To ensure missing data was not affecting our results, we projected each dataset down to 60%, 70%, and 80% of individuals present at each locus. Each model was optimized in at least five independent runs and convergence was assessed by achieving similar likelihood scores and parameter estimates. The best diffusion fit was used for comparing models using AIC scores to account for variable numbers of parameters estimated in each model. Since the models are nested, we conducted an adjusted likelihood-ratio test using the Godambe Information Matrix and 100 bootstrap replicates to check whether there is support for the model preferred by AIC (complex model). To do this, we evaluated the best-fit parameters from the simpler model using the complex model. Credible intervals were calculated using the Fisher Information Matrix (*FIM_uncert*). Finally, optimized parameters were converted to real time and population size units using a generation time of one year. Mutation rates are lineage-specific and scale with divergence time, therefore we conducted

conversions using a range of mutation rates: a flycatcher germline specific rate of 2.3 x $10^{-9}$ m/s/y [60], a zebra finch lineage rate of 2.2 x $10^{-9}$ m/s/y [61], and the mean galliform mutation rate of 1.35 x $10^{-9}$ m/s/ y[62].

**Results**

We successfully constructed a reference genome for *M. alba* which was used to assemble ddRADseq loci. A total of 219 million quality filtered reads were aligned with 5x average coverage to 90% of the zebra finch genome. Reference mapping of 442.5 million ddRADseq reads resulted in 8.2 x $10^{4}$ unique loci across 246 individuals with an average coverage of 29.7 (BioProject PRJNA356768).

For each species group, we analyzed 18 different dataset combinations consisting of either pseudo-unlinked or unlinked SNPs, and varying levels of missing data at the locus and individual level. On average, controlling for linkage reduced each dataset by 70%. This SNP pruning has potential to either remove biased SNPs or, if linkage was not a problem, to remove relevant information from the analysis. However, linkage did not alter our population structure estimates or Mantel tests. Overall, filtering data based on missing data at the individual level had a larger impact than filtering data based at the locus level. As individuals with excess missing data were removed, population structure estimators were less likely to find a consistent result than when loci were removed. We present the majority-rule analyses (50% missing data) and "pseudo-unlinked" SNPs.

*Phylogenetic analyses*

The *Motacilla* relationships inferred in our time-calibrated species tree (Fig. 1c) were confirmed by genome-wide SNPs (Fig. 1b). Relationships were, for the most part, consistent with current wagtail taxonomy, but incongruent with the mtDNA tree (Fig. 1a).

*Mitochondrial relationships* - The mtDNA tree (Fig. 1a) recovered an "African clade" consisting of four Afrotropical/Malagasy species and an "Eurasian clade", which includes the Afrotropical *M. aguimp.* Consistent with previous mtDNA studies, both *M. flava* and *M. citreola* are polyphyletic within the Eurasian clade. *M. flava* is separated into a western clade (Z) and an eastern clade, which is further divided into two sub-clades (X, Y). Only clade Y contains a monophyletic subspecies grouping of *M. f. tschutschensis*. Each *M. flava* clade has a corresponding *M. citreola* clade: *M. c. citreola* is split between clades X and Y, and clade Z includes the southern *M. c. calcarata*. There is no structure consistent with subspecies designations in *M. alba* or *M. flava* clade X or Z.

*Nuclear relationships* – Both the RAxML gene tree analysis of the concatenated SNP data (Fig. 1b) and the SNAPP species tree analysis (Fig. 1d) support the African and Eurasian clades, with the latter divided into two primary clades notable for their plumage coloration: clade A includes only "black-and-white" (melanin-based) species, whereas clade B contains only species with green/yellow (carotenoid-based) plumages. These nuclear relationships are strikingly incongruent with mtDNA in 1) placing *M. aguimp* within the "black-and-white" clade, 2) supporting the monophyly of both *M. flava* and *M. citreola*, and 3) placing *M. bocagii* within the African clade. In the SNAPP tree, the Afrotropical mainland species (*M. capensis* and *M. clara*) form a clade, whereas the insular species (*M. flaviventris* and *M. bocagii*) form another. RAxML (Fig. 1b) found high support for splits among species but no support for within-species relationships. Accordingly, within *M. alba*, *M. flava* and *M. citreola*, no subspecies are monophyletic.

*Time-calibrated species tree* - The topology of the time-calibrated species tree (Fig. 1c) is identical to the SNAPP tree (Fig. 1d); moreover, it includes *M. maderaspatensis*, for which no SNP data are available. The African and Eurasian clades split towards the end of the Pliocene (~3 mya), with early and relatively widely spaced divergences from the early Pleistocene in the former clade, and an explosive radiation within < 0.5 million years (my) during the late Pleistocene.

*Population structure*

At the species level, *de novo* population estimation methods were able to distinguish recognized species, except *M. aguimp*, which was grouped with western *M. alba* by both DAPC and ADMIXTURE. Narrowing our focus to phenotypically distinct subspecies, the full dataset showed genetic structure discordant with phenotype in all species complexes. Instead, eastern and western populations ($K$=2) were resolved by both ADMIXTURE and *de novo* DAPC in *M. citreola* (2 subspecies, Fig. 2) and *M. flava* (13 subspecies, Fig. 3). Depending on which *M. alba* (9 subspecies) + *M. aguimp* dataset was analyzed in DAPC, there was support for either $K=2$ or $K=3$. ADMIXTURE consistently supported $K=2$ (Fig. 4c). For all three datasets, individual assignments were consistent across methods when $K=2$. Finally, we found no difference in population structure when rare variants were either considered separately or ignored.

Ordination plots from DAPC analyses with subspecies defined *a priori* provide additional resolution of population structure in *M. flava* (Fig. 3b) and *M. alba* (Fig. 4b). In both species, the x-axis of the ordination plots show a correlation between diagnosability and longitude. In *M. flava*, all subspecies clusters overlap, suggesting genetic distinctiveness is limited. We note our sampling of *M. f. tschutschensis* included individuals from the easternmost portion of its range. However, *M. f. tschutschensis* is distributed from Central Eurasia through the Kamchatka Peninsula and into Western Alaska with its geographic center close to that of *M. f. macronyx* and *M. f. taivana.* Therefore, its position in the center of the DAPC ordination plot conforms to the idea that this pattern tracks longitude, and it is unclear whether these clusters correspond to subspecies distinctions or merely IBD.

The ordination plot generated by assigning *M. alba* samples to subspecies is influenced by missing data. The less stringent the threshold, the more distinct *M. a. subpersonata* becomes. This trend is found in both pseudo-unlinked and unlinked datasets, with the unlinked presenting the most

extreme cases. However, it is remarkable that in three out of four cases, *M. a. subpersonata* is more distinct than *M. aguimp* and all other *M. alba* subspecies (Fig. 4b). As no other subspecies displayed this behavior, we removed *M. a. subpersonata* and recompiled datasets. This resulted in *M. aguimp* samples being tightly clustered and diagnosable in all analyses. The remaining *M. alba* subspecies clusters are overlapping and do not show clustering consistent with subspecies, but do show clustering on either side of the vertical axis that is consistent with our *de novo* $K$=2 clusters. Taken together with evidence that Bayesian clustering methods may not work well with small sample sizes [63], we consider both $K$=2 (without *M. a. subpersonata*) and $K$=3 (*M. a subpersonata* included as own population) population histories in the following analyses.

Population structure in *M. cinerea* was not consistent across methods. Whereas ADMIXTURE supported a single panmictic population, DAPC supported an east-west split ($K$=2) consistent with the other Eurasian wagtails. These contrasting findings may be explained by simulation and other empirical studies, which have demonstrated that Bayesian clustering methods fail to detect structure when genetic divergence is very low [64,65].

*Isolation by distance*

When dealing with genetic data from evenly spaced samples from a spatially structured organism, the expected behavior of PCA is to return clusters that are related to the geographic origin of each individual sample [49]. However, spatial autocorrelation may bias interpretation of PC analyses. While Mantel tests generally support a history of IBD in all species groups (p-value<0.05), p-values vary according to missing data.

As recent studies demonstrate that sampling design can strongly bias interpretations of Mantel tests [55], we further explored IBD using EEMS, a program that can distinguish whether support for IBD is the result of either geographically distant and differentiated populations or a continuous cline in genetic differentiation. Strong linear relationships between predicted and observed genetic

dissimilarities confirm that the EEMS model fit our data [56]. To assess support for a true barrier, we examined plots of dissimilarity between pairs of sampled demes for non-linearity and resulting effective migration map for a singular uniform barrier. EEMS strongly supports the existence of a barrier between eastern and western *M. citreola* (Fig. 2), consistent with the findings of population structure estimators.

In *M. cinerea*, EEMS finds no support for a barrier, consistent with population structure estimates from ADMIXTURE (*K*=1) but not DAPC (*K*=2). We consider *M. cinerea* a single panmictic population with a strong signal of IBD in downstream analyses. This is consistent with a previously described pattern of shallow clinal plumage variation seen across much of Eurasia [12]. We were unable to include samples from two of the three subspecies (small populations restricted to either the Azores and Madeira Islands) and can only conclude that there is support for a single panmictic mainland Eurasian *M. cinerea*.

We found conflicting support for IBD in *M. flava*. Despite a linear trend in predicted and observed genetic dissimilarities, supporting IBD, there is also strong support for the existence of two discontinuous barriers (Fig. 3c). We interpreted this to mean that while IBD exists across the whole *M. flava* range, there are areas where gene flow is not significant. For assignment of subspecies into these populations, see Figure 1d and Figure 3. Since these results differ from those of *de novo* adegenet and ADMIXTURE, we tested whether DAPC can distinguish between these groups when given individual assignments. First, we fixed *K=3* and ran *de novo* DAPC to see if we could find the same populations resolved by EEMS. Second, we defined EEMS populations and ran DAPC. As both methods overwhelmingly supported the three populations found by EEMS, we used these definitions in all remaining analyses.

*M. alba* also showed a linear trend and a patchy effective migration map, lending support for IBD. However, a closer look at posterior mean of effective migration rates demonstrates that some areas have a markedly higher rate of migration than average (Fig. 4c). EEMS largely supports the east-west division of *M. alba* and the distinctness of *subpersonata*. However, an inverted Y-shaped

migration surface splits eastern *M. alba* into northern (congruent with *M. a. lugens* and *M. a. ocularis*) and southern (congruent with *M. a. leucopsis* and *M. a. alboides*) populations (Fig. 4c). To explore the north-south division further, we implemented DAPC, but it was unable to distinguish between these populations. These results may stem from low sample sizes, as EEMS is more robust to biased sampling than PCA methods [56].

*Demographic history*

To narrow the number of possible 2D models, we first ran 1D models on each population separately. *M. citreola*, *M. flava*, and *M. alba* were all modeled using 2D models. (We were unable to optimize 3D *dadi* models due to a sparse site frequency spectrum, and therefore, conducted pairwise 2D model comparisons for *M. flava* [per comm. R. Gutenkunst].) All six models were optimized and at least five independent runs converged on similar parameter values. Within species, missing data did not affect demographic model selection and produced overlapping 95% credible interval of parameter estimates. Therefore, we chose to present the 80% threshold in both paper figures and 2D site frequency spectra.

All species support the *split + growth + symmetric migration* model (Fig. 5a), where population size increases following divergence. All converted time estimates lie within the Pleistocene (0.14-1.8 million years ago; Fig. 5b). Time estimates largely overlap across *M. flava* and *M. citreola* populations, whereas time estimates for *M. alba* are more recent. Real time estimates are sensitive to the choice of multipliers, such as generation time and mutation rate. Therefore, we were primarily interested in relative population size and relative time estimates. However, all three avian mutation rates recover time estimates within the Pleistocene.

Finally, *M. cinerea* consists of a single panmictic population supporting a two-epoch model, where at time *T* the ancestral population underwent instantaneous population growth.

## Discussion

Modern phylogeographic studies assume that the population divergence history of widespread and variable species can be disentangled with sufficient numbers of variable, neutrally evolving SNPs [66,67]. We collected genome-wide SNPs to compare populations at different stages of the speciation continuum. Our data resolved species-level relationships with high support, suggesting that some populations are structured along an east-west axis, but failed to distinguish between recently diverged, phenotypically distinct taxa that are usually treated as subspecies. Recent differentiation, ongoing gene flow, and demographic processes may have obscured patterns of differentiation at neutral loci (cf. [68,69]).

### *Phylogeny*

We provide evidence that the currently recognized wagtail species are monophyletic and resolve the relationships among them. Within the "African" clade, we find a sister relationship between mainland and island species. Within the latter clade, *M. bocagii* is endemic to São Tomé, an island off the west coast of Africa, while *M. flaviventris* is endemic to Madagascar. Originally thought to belong to the sylvioid superfamily, *M. bocagii* bears little resemblance to other *Motacilla* species in both plumage, structural morphology, habitat, and behavior [21]. Our SNP data confirms its placement within *Motacilla*.

Species belonging to the "black-and-white" clade have mostly allopatric distributions, together covering nearly all of Europe, Asia, and Africa. While our *de novo* population structure analyses are unable to differentiate between *M. aguimp* and *M. alba, M. aguimp* is identifiable using *a priori* assignment. Moreover, our species tree analyses (Fig. 1d & Fig S#), which account for stochastic processes such as incomplete lineage sorting, found *M. aguimp* sister to *M. alba* with high

support, rather than embedded within *M. alba*. The distinctness of *M. aguimp* is further substantiated by its multiple differences (size, plumage, song, call) from all subspecies of *M. alba* [12].

We were unable to confirm the exact placement of species sister to the clade containing *M. alba* and *M. aguimp* due to the lack of *M. maderaspatensis*, a large-bodied resident of the Indian continent, in our SNP dataset. However, we confirmed monophyly of two of its members, *M. grandis* and *M. samveasnae*, which had been insufficiently sampled in previous studies. Endemic to Japan, *M. grandis* is a monotypic, short-distance migrant that is easily distinguishable from its partially sympatric relative *M. alba* by its plumage, larger body size, song and call [12]. Despite overlapping distributions and documented hybridization, no prior population genetic analyses of *M. alba* have included more than a single sample of *M. grandis* [17,70] and many have nested *M. grandis* within *M. alba* [12,17]. Using four newly sampled *M. grandis* individuals, our results show that *M. grandis* is distinct from *M. alba*. The close relative, *M. samveasnae*, has a comparatively minute distribution and is endemic to the Lower Mekong basin of Cambodia and Laos. In appearance, it closely resembles the Afrotropical *M. aguimp* but our increased sampling confirms its distinctness.

Previous mtDNA-based studies on *M. alba* delineated three or four populations. The only population consistent with our SNP results was the rare Moroccan endemic *M. a. subpersonata*, which was found to be divergent in both studies. Li *et al*. (2016) included denser sampling of the southeastern part of the *M. alba* range, but this sampling difference is unlikely to explain the incongruent population structure and instead reflects inherent differences between mitochondrial and nuclear data, such as mutational rate and inheritance patterns. Intra-specific relationships remain unresolved and divergence time estimates support a scenario of recent, rapid radiation.

Within the yellow plumage clade, this same trend holds true for both *M. flava* and *M. citreola*. *M. flava* and *M. citreola* are monophyletic and sisters, contrary to mtDNA findings and accordingly does not support Pavlova *et al*.'s proposal to split *M. citreola* into two species.

*Plumage divergence*

Discordance between phenotype and genotype, along with our divergence time estimates, suggest that wagtail plumage evolution has been very recent and rapid. This is particularly true for the Palearctic *M. flava* and *M. alba* complexes, which display exceptionally high plumage differentiation during the second half of the Pleistocene (as also suggested by [17] for *M. alba*). This has led to some remarkable cases of parallel evolution, e.g. in the widely disjunct *M. aguimp* vs. *M. samveasnae*; *M. f. thunbergi* vs. *M. f. macronyx*; and *M. f. flava* vs. *M. f. tschutschensis* (cf. Fig. 1). Moreover, the widely allopatric *M. f. flavissima*, *M. f. lutea* and *M. f. taivana* differ from the other *M. flava* subspecies by their bright yellow and green head patterns, and in the *M. alba* complex certain head pattern and upperpart coloration traits display a mosaic geographical pattern, indicating parallel evolution.

Most of the plumage similarities and differences among wagtail taxa concern differently coloured plumage patches. It seems possible that these can be switched on and off in different combinations through a rather simple system. Recent studies have demonstrated that strong selection can occur at few genes, and that plumage differences can evolve rapidly without corresponding divergence in the rest of the genome [10,72]. For example, within a Eurasian crow (*Corvus*) species complex, there is evidence of assortative mating based on plumage despite no genetic differentiation across most of the genome. Instead, there appears to be simple genetic control of phenotype, but divergence is occurring at different genes in different populations, thus implicating a more complex multi-genic pathway than previously thought [11]. While our SNP data was sampled from across the genome, it only represents 0.05% of the entire 1.1 Gb wagtail genome. If few genes are responsible for plumage differences and populations experienced recent divergence and selection, then it is not surprising that our data fails to find a genetic signal congruent with phenotype. Strong selection on plumage loci, mediated through assortative mating and selection against intermediate plumage phenotypes (hybrids), might explain how different plumage traits can be maintained in the face of gene flow in hybrid zones. Hybridization is prevalent, in variously wide hybrid zones, among

parapatrically distributed subspecies of *M. flava*, *M. citreola* and *M. alba*, and interspecific hybridizations, e.g. between *M. flava* and *M. citreola*, are not infrequent (reviewed in [12]).

Strong integrity of the *M. f. thunbergi* phenotype, but large-scale introgression of other loci from *M. f. tschutschensis* might perhaps explain the confusing pattern in the continuously distributed *M. f. thunbergi* (cf. Fig. 3), which displays a particularly strong genotype vs. phenotype discordance. Such a pattern has been found in the *Corvus corone* complex in western Europe, where adjacent populations of the phenotypically different *C. c. corone* and *C. c. cornix* are genetically more similar than some phenotypically similar populations of the former [72]. Alternatively, the *M. f. thunbergi* phenotype might have 'invaded' the *M. f. tschutschensis* genome through a selective sweep. Introgression of specific morphological traits have been noted in e.g. *Heliconius* butterflies (Pardo-Diaz et al. 2012, The *Heliconius* Genome Consortium 2012) and domestic chicken (Eriksson *et al*. 2008). However, neither of these explanations seems fully compatible with the genetic differentiation and gene flow restrictions between western and eastern *M. f. thunbergi* first detected by mtDNA by Pavlova et al. [16] and confirmed by SNP data in the present study. Populations east of c. 83–85°E are often recognized as *M. f. plexa* (e.g. [Red'kin & Babenko 1999]).

In general, *Motacilla* species vary considerably more in plumage traits closely linked to signalling functions than to ecological (e.g. beak, tarsus, or wing length) traits (cf. [74]), indicating that sexual selection might have played an important role in phenotypic divergence. Furthermore, the Eurasian clade displays variation in sexual dimorphism, size, seasonal plumage variation, and age-related differences, whereas the African clade does not [12,18], indicating greater potential for sexual selection in the former group. However, field experiments are necessary to confirm the role of plumage in mate preference. Sexually selected plumage traits have potential to diverge rapidly among allopatric, ecologically equivalent populations [1,73], and have been identified as the dominant signals mediating mate-choice and intrasexual aggression in some birds [74]. In contrast, the aberrant structure and plumage of *M. bocagii* (cf. Fig. 1) has been suggested to be the result of natural selection following colonization of a novel habitat [21].

The African clade is more than twice as old as the Eurasian clade, but contains only c. ¼ of the number of taxa (Gill & Donsker 2017), and much less plumage variation [12,18]. However, it includes both species with entirely melanin-based (black, brown) plumage and species with carotenoid-based (yellow, green) plumage patterns. In contrast, within the Eurasian clade, subclade A contains only species with 'black-and-white' (melanin-based) plumage, whereas the species in subclade B have strong carotenoid-based plumage patterns. Based on the distribution of yellow plumage in the different species (clade B, *M. flaviventris*, and to a lesser extent *M. capensis*), we hypothesize that *Motacilla*'s common ancestor had the ability to produce carotenoid-based colour patterns, whereas this pathway has been suppressed in clade A and the two other species in the African clade.

Further studies using transcriptomics...will hopefully shed light on the complex plumage evolution in wagtails

*Demographic history*

Many bird species have experienced range shifts in the Palearctic since the late Miocene [23,74,75]. However, evidence of avian divergence during the Pleistocene varies [24,76,77]. Here, we explicitly modelled patterns of population divergence, population size changes over time, and patterns of gene flow, placing Eurasian wagtail diversification occurring during the Pleistocene. Previous divergence estimates, based solely on mtDNA gene tree calibration, are much earlier [15,21], but are likely biased by multiple missing taxa.

If the pattern of intraspecific divergence in the *M. alba* and *M. flava* species complexes was caused by fragmentation of breeding ranges and prolonged periods of divergence in isolation, then we would expect to see corresponding genetic differentiation among subspecies and demographic models supporting deeper divergence times. However, if multiple interglacials allowed populations to repeatedly meet and mix, differentiation evolved in isolation could have eroded. Under this scenario,

it also possible that differential selection on plumage controlled by few genes could lead to the plumage diversity we see today. Alternatively, if subspecies are the result of recent and rapid selection following the end of the last glacial maxima, we would expect to see population structure incongruent with phenotype and a recent history of population expansion in demographic models.

While neither scenario is consistent with our *M. flava* or *M. citreola* results, we do find support for the latter in *M. alba* which bears the mark of the most recent climatic events. In *M. alba*, the 95% credible interval of population division (Fig. 5b) coincides with the start of a cooling period (~75,000 ybp; [78], which lasted until the last glacial maximum (LGM: ~22,000 ybp; [79]. Following the LGM, ice sheets receded and our estimates find a corresponding population increase during that time. These findings are consistent with past work on the *M. alba* species complex which also found evidence for Pleistocene population expansion and relatively recent introgression at the edge of their range [17,70]. Furthermore, the upper 95% credible interval divergence time estimates (Fig. 1c) predate those of our population divergence time estimates (Fig. 5b) and are therefore consistent.

Coincident patterns of barriers to gene flow in *M. flava* and *M. citreola* suggest that major eastern and western refugia existed. The western barrier coincides with a vast area of lowland (Turgai depression) that periodically became a huge body of water, potentially acting as a west-east barrier during interglacial periods [80]. Therefore, western and eastern populations may have been separated during both glacial and interglacial periods. Shorter periods of separation in minor refugia may have then led to plumage differentiation under strong sexual selection, but without more large-scale genetic divergence. Further separation of *M. flava* into northern and southern populations in the east is likely a result of East Asia being much less affected by glaciation. Indeed, other birds show similar divergence patterns as wagtails [6,81,82], indicating a common cause. For a more detailed reconstruction of complex demographic events, increased genomic sampling is needed [83].

While still supporting a history of Pleistocene diversification, both *M. citreola* and *M. flava* divergence time estimates (Fig. 1c) postdate those of demographic models. This inconsistency could be due to two factors. First, the *M. alba* specific mtDNA rate used to calibrate our species tree was

derived from the generic avian *cyt-b* rate. It is well established that molecular rates vary across lineages and this rate may be inaccurate. Second, gene flow leads to underestimation of divergence times in time-calibrated phylogenies [43]. Hybridization between *flava* and *citreola* has been documented numerous times (reviewed in [12]), predominantly on the expanding western edge of the *citreola* range, and may explain why divergence dates are more recent than our estimates population split time.

Once thought to be a panacea for resolving relationships between rapidly evolving lineages, genome-wide SNPs were unable to distinguish between phenotypically-distinct wagtail populations. It is clear that subspecies should not be treated as evolutionary units. Future studies should harness the power of whole-genome re-sequencing and gene expression studies, as changes in gene expression often underlie changes in phenotypic differentiation [84].

**Data Accessibility**

- Raw, demultiplexed ddRAD reads: NBCI SRA under PRJNA356768
- SNP datasets and input files for analyses: Dryad repository doi:10.5061/dryad.008bq

**References**


1. Price T. 2008 *Speciation in Birds*. Roberts & Company.

2. Mason NA, Taylor SA. 2015 Differentially expressed genes match bill morphology and plumage despite largely undifferentiated genomes in a Holarctic songbird. *Mol. Ecol.* **24**, 3009–3025.

3. Knowles LL, Richards CL. 2005 Importance of genetic drift during Pleistocene divergence as revealed by analyses of genomic variation. *Mol. Ecol.* **14**, 4023–4032.

4. Nosil P, Funk DJ, Ortiz-Barrientos D. 2009 Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**, 375–402.

5. Nosil P, Harmon LJ, Seehausen O. 2009 Ecological explanations for (incomplete) speciation. *Trends Ecol. Evol.* **24**, 145–156.

6. Olsson U, Alström P, Svensson L, Aliabadian M, Sundberg P. 2010 The Lanius excubitor (Aves, Passeriformes) conundrum—taxonomic dilemma when molecular and non-molecular data tell different stories. *Mol. Phylogenet. Evol.* **55**, 347–357.

7. Omland KE, Lanyon SM. 2000 Reconstructing plumage evolution in orioles (Icterus): repeated convergence and reversal in patterns. *Evolution* **54**, 2119–2133.

8. Milá B, McCormack JE, Castañeda G, Wayne RK, Smith TB. 2007 Recent postglacial range expansion drives the rapid diversification of a songbird lineage in the genus Junco. *Proc. Biol. Sci.* **274**, 2653–2660.

9. Poelstra JW, Vijay N, Hoeppner MP, Wolf JBW. 2015 Transcriptomics of colour patterning and coloration shifts in crows. *Mol. Ecol.* **24**, 4617–4628.

10. Toews DPL, Taylor SA, Vallender R, Brelsford A, Butcher BG, Messer PW, Lovette IJ. 2016 Plumage genes and little else distinguish the genomes of hybridizing warblers. *Curr. Biol.* **26**,

2313–2318.

11. Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, Wolf JBW. 2016 Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nat. Commun.* **7**, 13195.

12. Alström P, Mild K. 2003 *Pipits and Wagtails of Europe, Asia and North America*. A&C Black, Princeton University Press.

13. Pavlova A, Zink RM, Rohwer S, Koblik EA, Red'kin YA, Fadeev IV, Nesterov EV. 2005 Mitochondrial DNA and plumage evolution in the white wagtail *Motacilla alba*. *J. Avian Biol.* **36**, 322–336.

14. Ödeen A, Björklund M. 2003 Dynamics in the evolution of sexual traits: losses and gains, radiation and convergence in yellow wagtails (*Motacilla flava*). *Mol. Ecol.* **12**, 2113–2130.

15. Voelker G. 2002 Systematics and historical biogeography of wagtails: dispersal versus vicariance revisited. *Condor* **104**, 725.

16. Pavlova A, Zink RM, Drovetski SV, Red'kin Y, Rohwer S. 2003 Phylogeographic patterns in *Motacilla flava* and *Motacilla citreola*: species limits and population history. *Auk* **120**, 744.

17. Li X *et al.* 2016 Shaped by uneven Pleistocene climate: mitochondrial phylogeographic pattern and population history of white wagtail *Motacilla alba* (Aves: Passeriformes). *J. Avian Biol.* **47**, 263–274.

18. del Hoyo J, Elliott A, Sargatal J. 2004 *Handbook of the Birds of the World: Cotingas to pipits and wagtails*.

19. Alström P, Ödeen A. 2002 Incongruence between mitochondrial DNA , nuclear DNA and non-molecular data in the avian genus *Motacilla*: implications for estimates of species

phylogenies.

20. Ödeen A. 2001 Effects of post-glacial range expansions and population bottlenecks on species richness.

21. Alström P, Jønsson KA, Fjeldså J, Ödeen A, Ericson PGP, Irestedt M. 2015 Dramatic niche shifts and morphological change in two insular bird species. *Royal Society Open Science* **2**, 140364–140364.

22. Hewitt GM. 2000 The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913.

23. Voelker G. 2010 Repeated vicariance of Eurasian songbird lineages since the Late Miocene. *J. Biogeogr.* **37**, 1251–1261.

24. Nadachowska-Brzyska K, Li C, Smeds L, Zhang G, Ellegren H. 2015 Temporal dynamics of avian populations during Pleistocene revealed by whole-genome sequences. *Curr. Biol.* **25**, 1375–1380.

25. Hewitt GM. 2004 Genetic consequences of climatic oscillations in the Quaternary. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **359**, 183–195.

26. Martin PR, Montgomerie R, Lougheed SC. 2010 Rapid sympatry explains greater color pattern divergence in high latitude birds. *Evolution* **64**, 336–347.

27. Stewart JR, Lister AM, Barnes I, Dalén L. 2010 Refugia revisited: individualistic responses of species in space and time. *Proc. Biol. Sci.* **277**, 661–671.

28. Gohli J, Leder EH, Garcia-Del-Rey E, Johannessen LE, Johnsen A, Laskemoen T, Popp M, Lifjeld JT. 2015 The evolutionary history of Afrocanarian blue tits inferred from genomewide SNPs. *Mol. Ecol.* **24**, 180–191.

29. Rokas A, Carroll SB. 2006 Bushes in the tree of life. *PLoS Biol.* **4**, e352.

30. 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.

31. Dray S, Josse J. 2014 Principal component analysis with missing values: a comparative survey of methods. *Plant Ecol.* **216**, 657–667.

32. Huang H, Knowles LL. 2014 Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Syst. Biol.* **65**, 357–365.

33. Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013 Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* **22**, 787–798.

34. Nakagawa S, Freckleton RP. 2008 Missing inaction: the dangers of ignoring missing data. *Trends Ecol. Evol.* **23**, 592–596.

35. Maddison WP. 1997 Gene trees in species trees. *Syst. Biol.* **46**, 523.

36. Drummond AJ, Rambaut A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214.

37. Heled J, Drummond AJ. 2012 Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.* **61**, 138–149.

38. Ritchie AM, Lo N, Ho SYW. 2016 The Impact of the Tree Prior on Molecular Dating of Data Sets Containing a Mixture of Inter- and Intraspecies Sampling. *Syst. Biol.* (doi:10.1093/sysbio/syw095)

39. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large

phylogenies. *Bioinformatics* **30**, 1312–1313.

40. Lischer HEL, Excoffier L, Heckel G. 2014 Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of microtus voles. *Mol. Biol. Evol.* **31**, 817–831.

41. Leaché AD, Banbury BL, Felsenstein J, de Oca AN-M, Stamatakis A. 2015 Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Syst. Biol.* **64**, 1032–1047.

42. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012 Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29**, 1917–1932.

43. Leaché AD, Harris RB, Rannala B, Yang Z. 2013 The influence of gene flow on species tree estimation: a simulation study. *Syst. Biol.* **63**, 17–30.

44. Ödeen A, Björklund M. 2003 Dynamics in the evolution of sexual traits: losses and gains, radiation and convergence in yellow wagtails (*Motacilla flava*). *Mol. Ecol.* **12**, 2113–2130.

45. Pavlova A, Zink RM, Drovetski SV, Red'kin Y, Rohwer S. 2003 Phylogeographic patterns in *Motacilla flava* and *Motacilla citreola*: species limits and population history. *Auk* **120**, 744.

46. Jombart T, Devillard S, Balloux F. 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.

47. Jombart T, Devillard S, Dufour A-B, Pontier D. 2008 Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**, 92–103.

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

49. Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, Clark AG. 2010 Inferring genetic ancestry: opportunities, challenges, and implications. *Am. J. Hum. Genet.* **86**, 661–673.

50. Gompert Z, Lucas LK, Alex Buerkle C, Forister ML, Fordyce JA, Nice CC. 2014 Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Mol. Ecol.* **23**, 4555–4573.

51. Wright S. 1943 Isolation by Distance. *Genetics* **28**, 114–138.

52. Dray S, Dufour A-B. 2007 The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**. (doi:10.18637/jss.v022.i04)

53. Bivand RS, Pebesma E, Gómez-Rubio V. 2013 *Applied spatial data analysis with R*.

54. Legendre P, Fortin M-J, Borcard D. 2015 Should the Mantel test be used in spatial analysis? *Methods Ecol. Evol.* **6**, 1239–1247.

55. Guillot G, Rousset F. 2013 Dismantling the Mantel tests. *Methods Ecol. Evol.* **4**, 336–344.

56. Petkova D, Novembre J, Stephens M. 2016 Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100.

57. BirdLife International and NatureServe. 2015 Bird species distribution maps of the world.

58. Plummer M, Best N, Cowles K, Vines K. 2006 CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11.

59. 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.

60. Smeds L, Qvarnström A, Ellegren H. 2016 Direct estimate of the rate of germline mutation in a bird. *Genome Res.* **26**, 1211–1218.

61. Nam K *et al.* 2010 Molecular evolution of genes in avian genomes. *Genome Biol.* **11**, R68.

62. Ellegren H. 2007 Molecular evolutionary genomics of birds. *Cytogenet. Genome Res.* **117**, 120–130.

63. Fogelqvist J, Niittyvuopio A, Agren J, Savolainen O, Lascoux M. 2010 Cryptic population genetic structure: the number of inferred clusters depends on sample size. *Mol. Ecol. Resour.* **10**, 314–323.

64. Waples RS, Gaggiotti O. 2006 What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol. Ecol.* **15**, 1419–1439.

65. Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L. 2015 RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Mol. Ecol.* **24**, 3299–3315.

66. Brumfield RT, Beerli P, Nickerson DA, Edwards SV. 2003 The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol.* **18**, 249–256.

67. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013 Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* **66**, 526–538.

68. Campagna L, Benites P, Lougheed SC, Lijtmaer DA, Di Giacomo AS, Eaton MD, Tubaro PL. 2011 Rapid phenotypic evolution during incipient speciation in a continental avian radiation.

*Proceedings of the Royal Society B: Biological Sciences* **279**, 1847–1856.

69. André C *et al.* 2010 Detecting population structure in a high gene-flow species, Atlantic herring (*Clupea harengus*): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity* **106**, 270–280.

70. Pavlova A, Zink RM, Rohwer S, Koblik EA, Red'kin YA, Fadeev IV, Nesterov EV. 2005 Mitochondrial DNA and plumage evolution in the white wagtail *Motacilla alba*. *J. Avian Biol.* **36**, 322–336.

71. Poelstra JW *et al.* 2014 The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**, 1410–1414.

72. Price T. 1998 Sexual selection and natural selection in bird speciation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **353**, 251–260.

73. Seddon N *et al.* 2013 Sexual selection accelerates signal evolution during speciation in birds. *Proc. Biol. Sci.* **280**, 20131065.

74. Holm SR, Svenning J-C. 2014 180,000 years of climate change in Europe: avifaunal responses and vegetation implications. *PLoS One* **9**, e94021.

75. Drovetski SV. 2003 Plio-Pleistocene climatic oscillations, Holarctic biogeography and speciation in an avian subfamily. *J. Biogeogr.* **30**, 1173–1181.

76. Zink RM, Slowinski JB. 1995 Evidence from molecular systematics for decreased avian diversification in the Pleistocene epoch. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 5832–5835.

77. Klicka J. 1997 The Importance of Recent Ice Ages in Speciation: A Failed Paradigm. *Science* **277**, 1666–1669.

78. Ehlers J, Gibbard PL, Hughes PD. 2011 *Quaternary glaciations - extent and chronology: a*

*closer look*. Elsevier.

79. Ehlers J, Gibbard P. 2008 Extent and chronology of Quaternary glaciation. *Episodes* **31**, 211–218.

80. Zubakov VA, Borzenkova II. 1990 *Global Palaeoclimate of the Late Cenozoic*. Elsevier.

81. Zink RM, Pavlova A, Drovetski S, Rohwer S. 2008 Mitochondrial phylogeographies of five widespread Eurasian bird species. *J. Ornithol.* **149**, 399–413.

82. Olsson U, Leader PJ, Carey GJ, Khan AA, Svensson L, Alström P. 2013 New insights into the intricate taxonomy and phylogeny of the *Sylvia curruca* complex. *Mol. Phylogenet. Evol.* **67**, 72–85.

83. Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. 2016 PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Mol. Ecol.* **25**, 1058–1072.

84. Brawand D *et al.* 2011 The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348.

Figure 1. (a) BEAST mitochondrial gene tree and (b) RAxML consensus tree of 500 random haplotype datasets; in both, nodes of monophyletic, single-species clades are collapsed for ease of viewing. Numbers in parentheses indicate sample size. (c) *BEAST tree inferred from mtDNA and nuclear introns, calibrated using a 2.7% *ND2* rate. Node bars indicate the 95% HPD of height. (d) SNAPP maximum clade credibility species tree (1467 biallelic SNPs with 8.5% missing data) rooted with *Dendronanthus* (not shown). In all trees, posterior probability is indicated at the nodes, where a circle denotes PP≥0.95. In all figures, paintings by Bill Zetterström and Ren Hathway.

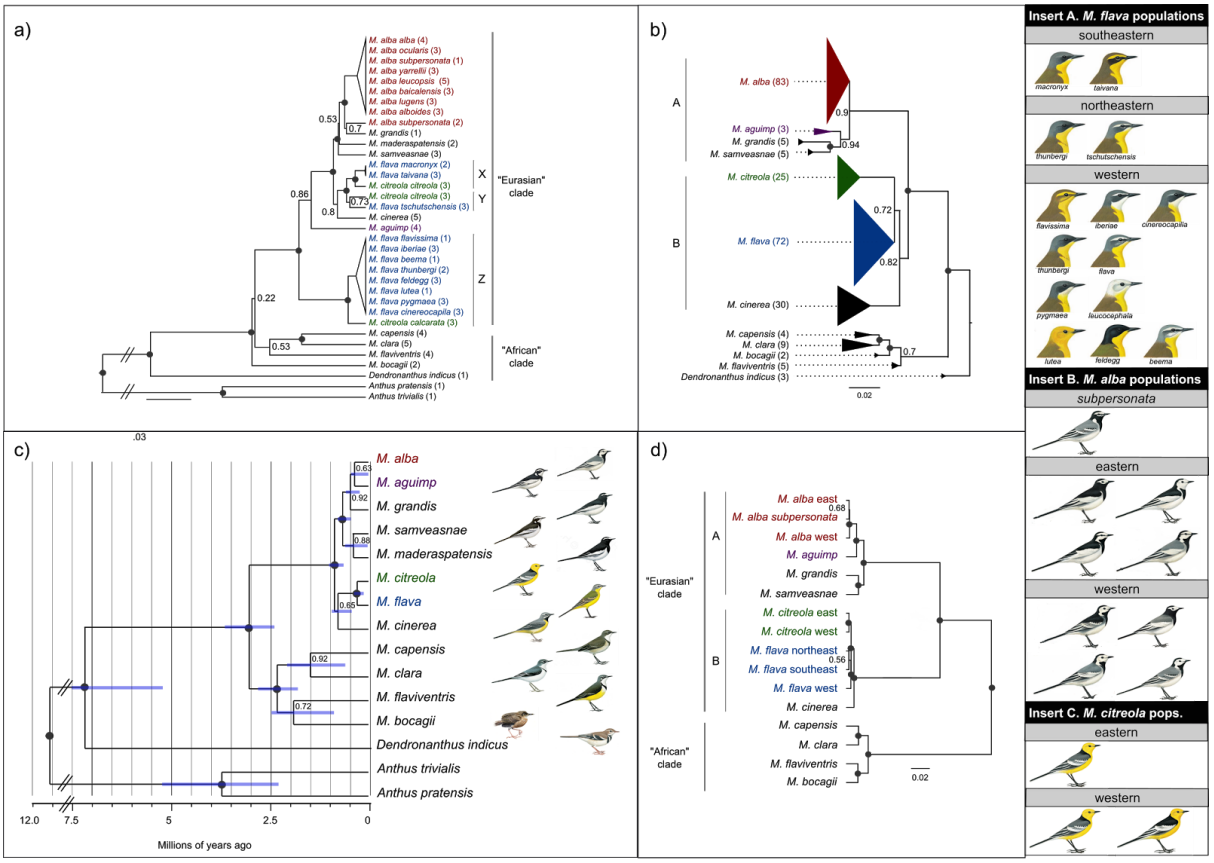


Figure 2. Posterior probabilities of effective migration rates of *M. citreola* estimated by EEMS. Birds sampled in India belong to *M. c. calcarata*, all others to *M. c. citreola*. In all figures, pie charts are located at sampling sites and denote the posterior probability of ADMIXTURE assignments (here, *K*=2).

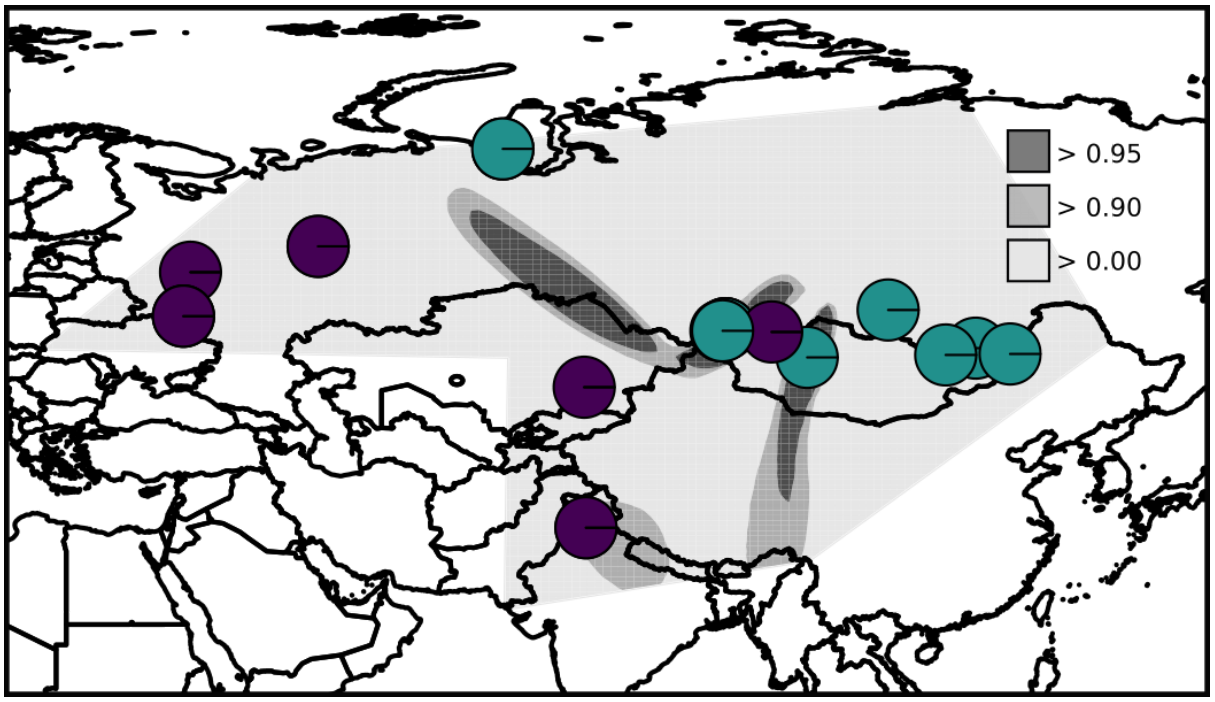


Figure 3. (a) *M. flava* breeding distribution and sampling localities. (b) DAPC plot of genetic clustering by subspecies (c) Posterior probability of effective migration rates estimated by EEMS show two barriers and *K*=3.

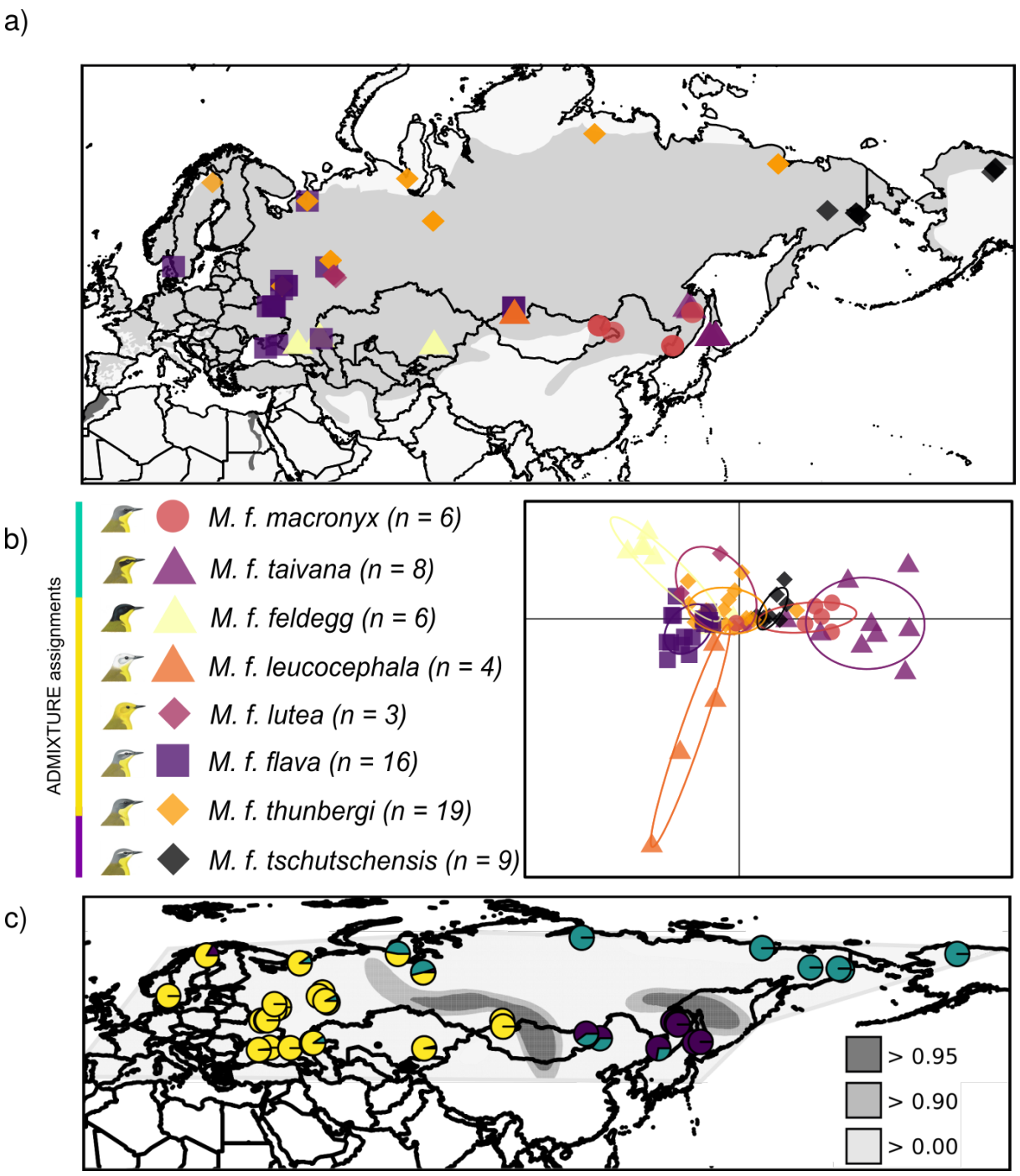


Figure 4. (a) *M. alba* distribution and sampling localities. (b) DAPC plot of genetic clustering by subspecies. (c) Posterior probabilities of estimated effective migration rates show 2-3 barriers and *K*=2.

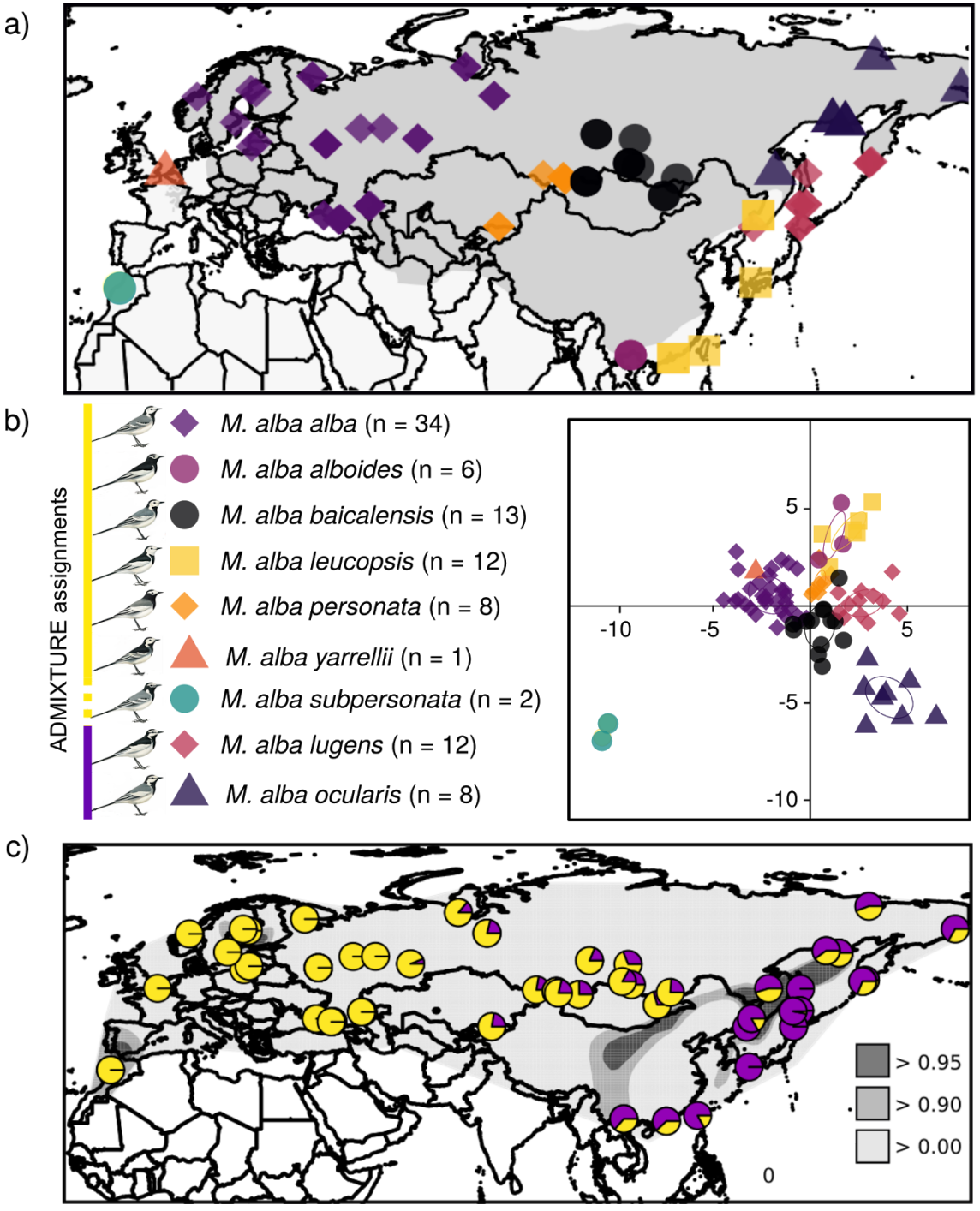
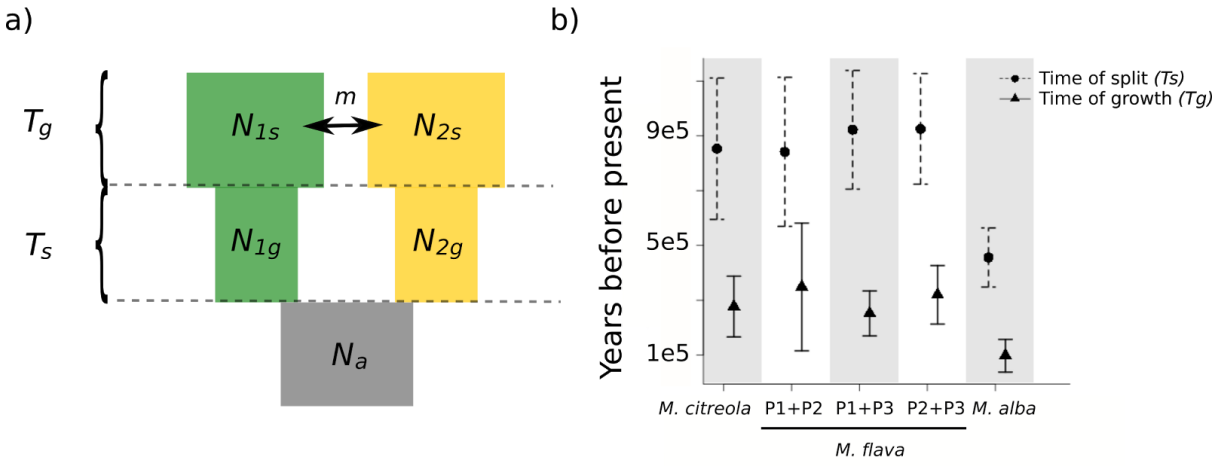


Figure 5. (a) Schematic representation of the best-fit demographic model, *split + growth + symmetric migration*. (b) Comparison of 2D *dadi* time estimates.



**Chapter 2:** A reference genome for *Motacilla flava*

Rebecca B. Harris[1,2] and Per Alström[3,4,5]

[1] Department of Biology, University of Washington, Seattle, WA 98195

[2] Burke Museum of Natural History and Culture, Seattle, WA 98195

[3] Department of Animal Ecology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[4] Swedish Species Information Centre, Swedish University of Agricultural Sciences, Box 7007, Uppsala SE-750 07, Sweden

[5] Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

## Introduction

Reference genomes are needed to understand the broad-scale processes and patterns important to genome evolution. In contrast to genetic sequences from individual loci, a genome reveals how genetic material is organized and provides a more complete picture of sequence evolution. Data on base composition, abundance of genes, or recombination landscapes helps researchers understand the biology of a given genome and infer factors that are likely to have played a role in its evolution [1]. For example, recombination directly affects selection when selection at linked sites interferes with selection at a focal site. Understanding how these two mechanisms interact necessitates the study of the genome as a whole.

Recent genome sequencing of nonmodel organisms has led to the identification of lineage-specific genes likely underlying phenotypic evolutionary novelty [2,3]. At the population level, whole-genomes provide a platform for analyses of genome wide polymorphism data and the study of genomic landscape and variance of allelic diversity within and between populations [4]. Genome scans for regions of high divergence between populations implicate reduced introgression and the role of diversifying selection. Because these studies rely on the visualization of peaks of divergence (frequently measured with Fst) along chromosomes -- and often use sliding windows to aggregate information from surrounding loci -- they require the information about genome order and composition provided by a reference genome [5].

The development of reduced-representation library methods, like ddRAD (Chapter 1), has been praised as an efficient method for sampling genome-wide SNPs for phylogenetic studies [6] and constructing linkage maps [7] of non-model organisms with no prior genomic information. However, systematic genotyping errors may lead to spurious linkages and incorrect ordering [8]. Furthermore, the clustering of ddRAD reads into loci can be adversely affected when a genome is highly heterozygous or repetitive. *De novo* clustering methods employ a variety of user-input thresholds to determine how loci are clustered, but researchers often lack the knowledge to make informed

parameter choices, despite the potential for biased values to have downstream effects that impact biological interpretation. Similarly, the widespread assumption in population genetic analyses that loci are unlinked is rarely tested. Finally, RAD sequencing provides only a sparse sampling of loci from across the genome. Reference genomes can alleviate these problems by providing information on composition of loci, their chromosomal location, and linkage patterns.

Compared to other vertebrates, birds have a relatively constrained genome size, stable chromosomal structure, and conserved gene order [9,10]. However, neutral and adaptive evolutionary processes have shaped avian genomes through lineage-specific substitution rates, with Passeriformes exhibiting the highest evolutionary rate [9]. Alignment of short (< 50 bp) single-end reads to distant relatives may be difficult with elevated rates. Though other passerine genomes have been sequenced, we were unable to align most wagtail ddRAD reads (Chapter 1) to available species. To resolve issues surrounding ddRAD read clustering (Chapter 1) and facilitate exploration of the evolution of phenotypic traits (Chapter 3), we assembled and annotated a reference genome for *M. flava tschutschensis*. Our choice was guided by a previous study suggesting postglacial founder events and population bottlenecks driving phenotypic differentiation in the northern *M. flava* populations [12]

## Methods

### *Sequencing*

Heterozygous sites may have an adverse effect on genome assembly [11]. Ideally, a reference genome should be created from an inbred line. However, we were limited to samples of wild birds accessioned in museum collections. Liver tissue from a single *M. flava tschutschensis* was sent to the Swedish National Genome Infrastructure for library preparation and sequencing. A short-insert library (180 bp) and two long-insert libraries (3 kbp and 6 kbp) were prepared using a Nextera-Mate Pair Kit

and sequenced in Rapid High Output mode on two lanes of Illumina HiSeq 2500 with paired-end 125bp reads.

*De novo assembly*

Currently, there is no optimal *de novo* assembler and performance can vary markedly when using same assembler method on slightly different data sets [11]. For these reasons, we did not limit our *de novo* analysis to a single tool, instead we employed three of the most commonly used assemblers: abyss [13], ALLPATHS-LG [14], and SOAPdenovo2 [15]. The ALLPATHS-LG method performs trimming and error checking from within the pipeline and requires raw reads as input. For the other two methods, we input quality filtered reads. Reads were trimmed of Illumina adapter sequences using Trimmomatic [16] and quality was checked using FastQC. To assess the complexity of our library in absence of a reference sequence, we then examined the kmer profile of the reads.

To determine the best assembler for our dataset, we conducted post-assembly evaluation using standard assembly statistics, coverage plots, and the feature-response curve (FRCurve). To avoid problems with outlier points and circumvent inherent differences in how assembler methods trim short contigs, we only used contigs longer than 1000bp during validation. All summary statistics were computed on the expected genome length in order to normalise the numbers for comparison among methods. Coverage information and FRCurve features were obtained by aligning the same reads used in the assembling phase against the assembled sequences using BWA-MEM algorithm [17].

Finally, we used BUSCO [18] to evaluate the completeness of each assembly in terms of gene content. We used the eukaryotic specific dataset of single-copy orthologous genes to query each assembly. Genes that were recovered from the assembly were classified as either: 1) complete = falls within two standard deviations of the expected gene length, 2) fragmented = if it does not fall within two standard deviations, or 3) duplicated = if two complete copies of a gene are recovered.

*Assembly validation and improvement*

*Checking mate-pair insert size* - Genome assembly methods assume sequencing reads are of known orientation and insert size, however poor library preparation can lead to mate-pair libraries contaminated with reads of unexpected orientation or insert size, thus introducing errors into genome assembly [19]. To confirm that mate-pair insert sizes were as expected, we mapped quality filtered mate-pair library reads back to our allpaths *de novo* assembly using Bowtie2 [20]. We set a wide range +/- 2 kbp. Using picard tools, we sorted the alignment by coordinate (*SortSam*) and then collected insert size metrics (*CollectInsertSizeMetrics*). We then checked the normality of the distribution of MP + PE fragment sizes using the CollectInsertSizeMetrics tool and determined whether their distributions were separated by visualizing histograms.

*Contamination*- We did not expect any contamination from outside sources. To confirm this, we searched for matching hits to our scaffolds using the BLAST nucleotide (nt) database and returned the five highest high-segment probability pairs with percent identity > 90% for each scaffold. If contamination were present, we would expect the whole genome of the contaminant to be present and not just short fragments. Therefore, short matches (< 5000 bp) were discarded. We then searched for the identity of these 5+ kbp matches using the NCBI GI number.

*Re-scaffolding* - Following the initial contig and scaffolding building step, we conducted a second scaffolding step to ensure proper genome order and orientation. Scaffolding utilizes the information from long-read libraries to join together contigs. One issue inherent to the scaffolding process is contamination of mate-pair libraries with paired-end reads. Paired-end contamination is a direct result of library preparation and can have adverse downstream effects because they are oriented differently and have shorter insert sizes the rest of the mate-pair library, leading to an increased number of misassemblies and inflated assembly sizes. Therefore, we implemented the stand-alone application BESST [21], a program that specifically models paired-end contamination. BESST

requires raw reads sorted according to orientation, trimmed of adaptors, and aligned to scaffolds. We implemented the mate-pair quality filtering with NxTrim [22]. Interlaced fastq data were then split into their respective files using bbmap and quality was checked in FastQC. Mate-pair were then aligned to scaffolds file using Bowtie2.

Whereas the goal of assembly and rescaffolding is to enhance contiguity, other methods use information from remapped reads to split scaffolds based on coverage and sequence composition. Regions with low coverage or mis-orientation of read pairs suggest poor assembly, whereas high sequence coverage may indicate the presence of collapsed, near identical repeats. We identified regions deviating from smooth, non-uniform coverage using Recognition in Errors using Aligned Paired Reads [23]. REAPR requires quality filtered, short-insert reads mapped back to the BESST scaffolds. Short-insert reads were quality filtered using Trimmomatic and mapped back to the BESST scaffolds using the -x option in smalt, as recommended in the REAPR documentation. REAPR then broke up misassembled scaffolds.

*Masking* - Repetitive regions can cause issues for genome alignment. To prevent against erroneous alignment in downstream analyses, we masked repetitive elements using RepeatMasker version open-4.0.6 [24] with the setting -species = “aves”.

*Validation* - To validate our *de novo* assembly, we aligned our masked *de novo M. flava* genome to the zebra finch genome. The zebra finch genome is the closest relative to *M. flava* that has a well-curated and anchored genome.The chromosomal arrangement of bird genomes are relatively conserved [10], making the percentage of scaffolds mapping to a well-curated bird genome a good metric for the completeness of our *de novo* genome. We used the fast alignment tool NUCmer [25] to find 500 bp or longer regions of exact match. NUCmer then extends its search algorithm outwards from the exact match to align genomes. These alignments were used to generate hypotheses about the arrangement, order, and anchoring of scaffolds to chromosomes.

*Genome Annotation*

We conducted *ab initio* genome annotation of the yellow wagtail genome using the gene prediction pipeline, MAKER [26]. For *ab initio* gene prediction, we utilized gene evidence from the *Ficedula* flycatcher. Specifically, we downloaded both the cDNA and protein databases from Ensembl and used these as evidence to predict genes in an iterative manner. In the first run, MAKER was run using gene evidence. Then, we trained the *ab initio* gene predictor SNAP using fathom and forge, and split annotations into four categories (unique genes, warnings, alternative spliced genes, and overlapping genes). These were exported and used to generate a HMM which was used in the second MAKER run.

To obtain biologically relevant protein descriptions, we first searched against the RefSeq database for the NCBI accession identifiers of high identity matches using *blastp*. These were then converted to relevant descriptions using the biomaRt package [27].

**Results**

*Raw reads*

Sequencing of *M. f. tschutschensis* resulted in 265.1 million read pairs with a PhiX error rate 0.505% and an average quality score of 34.63 (91.02% bases >= Q30). The kmer profile shows two peaks, indicating that the *M. f. tschutschensis* genome is highly heterozygous (Fig. 1). The second peak (centered at 57 bp) indicates the appropriate choice of kmer size in downstream analyses.

*De novo assembly*

*De novo* assembly can be assessed based on a number of factors: the length for which contigs of that length or longer contain at least 50% of the total of the lengths of all contigs (N50), total

genome size, coverage, level of contamination, biases in scaffold length, and the number of features. The average size of the passerine genome is 1.4 gb [28]. Evidence from karyotyping studies suggests that *M. flava* should be consistent with the average. Both ALLPATHS-LG and abyss produce assemblies within reason, whereas SOAPdenovo produced a very large and unlikely assembly (Table 1). ALLPATHS-LG had the longest N50.

The contig coverage distribution is expected to resemble as Gaussian distribution with the maximum around the expected coverage. Both ALLPATHS-LG and abyss display Gaussian-like distributions with peaks close to 80x coverage (Fig. 2-3), whereas SOAPdenovo does not (Fig. 4). Across all methods, there was no evidence of contamination - depicted by a random spread of GC content versus coverage points (Fig. 2-4). Both the GC content versus contig length, and the median coverage versus contig length graphs can reveal any biases in scaffold length. SOAPdenovo showed an increasing linear trend for median coverage versus contig length (Fig. 4), a trend not seen in abyss or ALLPATHS-LG. ALLPATHS-LG reconstructed a higher proportion of the genome with fewer features and higher coverage (Fig. 5).

ALLPATHS-LG assembly found 277 out of 429 complete eukaryotic genes, markedly more than the other two assemblers (Table 2). As a proxy for completeness, we compared this to the well-curated flycatcher genome and found 244 complete genes.

All evidence suggests that ALLPATHS-LG produced the best *de novo* assembly. A summary of the ALLPATHS-LG assembly can be found in Table 3. For the remainder of the manuscript, we refer to this assembly as our reference genome. All assembly validation and improvement was done using the ALLPATHS-LG assembly.

*Assembly validation and improvement*

Paired-end contamination can cause unexpected mate-pair insert sizes. Our mate-pair libraries insert sizes were no greater than 15% deviant from the expected insert size (Table 4), suggesting that

assembly was not adversely impacted by paired-end contamination. Nor was it impacted by contamination from outside sources. Our BLAST search found no repeat non-avian hits.

Rescaffolding with BESST increased the N50 of our assembly and decreased the total number of scaffolds (Table 1). This reduction was mostly due to long spanning reads joining two scaffolds and filling in the unsequenced regions with N's. However, this type of information is unnecessary for downstream analyses and REAPR broke these low coverage spans, increasing the number of scaffolds (Table 1).

A total of 3.63% of the *M. flava* genome was masked. Most of the identified repeat regions in *M. flava* were retro-elements which made up 3.34% of the assembly (Table 5). Repeat masking converted a total of 23 scaffolds into a string of N's and these were deleted from the working *M. flava* genome. The final *M. flava* consists of 18,512 scaffolds with 42.5% GC-content and an average depth of 67x.

*Genome completeness*

The final *M. f. tschutschensis* assembly covers nearly all (>90%) of the zebra finch genome (Fig. 6). We did not expect to find reads mapping to linkage group 2 or 5 as these are likely exclusive to the zebra finch and are absent from most other passerine genomes.

*Ab initio gene predictions*

Two rounds of *ab initio* gene prediction resulted in 9,463 genes with an average length of 20,753 bp and covering 18.7% of the *de novo* assembly (Table 6). The final gene models contain hits to 10,637 out of the 12,615 possible *Ficedula* proteins available on the UniProt database. The majority of proteins found in *M. flava* are either domain or transmembrane (Fig. 7).

**Discussion**

We present a first-pass reference genome of *M. flava* which will provide a stepping stone for a wide range of studies focused on the genus *Motacilla*. There is ample room for improvement of the reference including merging scaffolds and filling gaps and low coverage regions. A recent study consistently found ~15,000-16,000 genes across 48 bird species representing all extant clades [9], which suggests that the *M. flava* genome presented here is missing approximately 35% of the expected catalogue of avian genes. This is likely due to protein coding regions with low coverage or quality being broken up during the REAPR step of assembly validation.

The present genome will be vastly improved by the inclusion of additional mate-pair libraries with longer insert size or long read sequences (such as those from 10x, PacBio, or Hi-C). Such information will provide useful information on gene synteny. While birds are thought to have comparatively conserved genome arrangements, karyotype studies demonstrate that within *Motacilla* there have been a number of small rearrangements [29]. Furthermore, the proportion of repetitive elements (3.63%) is slightly less than that observed across birds (4-10%) [9]. This suggests that repetitive regions may have been collapsed, leading to the underestimation of repeat regions. However, many research questions do not necessitate a complete reference genome (i.e. near-complete DNA sequence for each chromosome).

The choice of *M. flava tschutshcensis* for genome sequencing was guided by the hypothesis that founder events and population bottlenecks led to reduced heterozygosity in *M. flava tschutshcensis* relative to other *M. flava* subspecies. However, we found no evidence for this demographic history (Chapter 1). Instead, *M. flava tschutshcensis* belongs to a northwestern subspecies group, along with *M. f. thunbergi*. Therefore, the heterogeneity observed in the kmer profile (Fig. 1) is an expected and unavoidable result, likely characteristic of all *M. flava* populations. Recent advancements in single-molecule sequencing methods will soon allow for phased genome

sequence reads [30], which will solve many of the issues surrounding the assembly of highly heterozygous genomes.

**References**


1. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29: 51–63.

2. Delmore KE, Toews DPL, Germain RR, Owens GL, Irwin DE. The Genetics of Seasonal Migration and Plumage Color. Curr Biol. 2016;26: 2167–2173.

3. McGaughran A, Rödelsperger C, Grimm DG, Meyer JM, Moreno E, Morgan K, et al. Genomic Profiles of Diversification and Genotype-Phenotype Association in Island Nematode Lineages. Mol Biol Evol. 2016;33: 2257–2272.

4. Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, et al. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat Commun. 2016;7: 13195.

5. Mascher M, Stein N. Genetic anchoring of whole-genome shotgun assemblies. Front Genet. 2014;5: 208.

6. Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, et al. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol Ecol. 2013;22: 787–798.

7. Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics. 2011;188: 799–808.

8. Henning F, Lee HJ, Franchini P, Meyer A. Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping.

Mol Ecol. 2014;23: 5224–5240.

9. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346: 1311–1320.

10. Ellegren H. Evolutionary stasis: the stable chromosomes of birds. Trends Ecol Evol. 2010;25: 283–291.

11. Ekblom R, Wolf JBW. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014;7: 1026–1042.

12. Odeen A, Björklund M. Dynamics in the evolution of sexual traits: losses and gains, radiation and convergence in yellow wagtails (Motacilla flava). Mol Ecol. 2003;12: 2113–2130.

13. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19: 1117–1123.

14. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008;18: 810–820.

15. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1: 18.

16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.

17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

18. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.

2015;31: 3210–3212.

19. Sahlin K, Chikhi R, Arvestad L. Assembly scaffolding with PE-contaminated mate-pair libraries. Bioinformatics. 2016;32: 1925–1932.

20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

21. Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L. BESST--efficient scaffolding of large fragmented assemblies. BMC Bioinformatics. 2014;15: 281.

22. O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics. 2015;31: 2035–2037.

23. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14: R47.

24. Smit A. Repeatmasker [Internet]. 2013-2015. Available: http://www.repeatmasker.org

25. Delcher AL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002;30: 2478–2483.

26. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18: 188–196.

27. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4: 1184–1191.

28. Website [Internet]. [cited 19 Apr 2017]. Available: Gregory, T.R. (2005). Animal Genome Size Database. http://www.genomesize.com.

29. Hammar B, Herlin M. Karyotypes of four bird species of the order Passeriformes. Hereditas. 1975;80: 177–184.

30. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49: 643–650.

Table 1. Standard summary statistics of the *M. flava* genome.

| Assembler | # Scfs | # Scfs > 1000 | NG50* | Max scf length | Total assembly length |
|---|---|---|---|---|---|
| abyss | 3611549 | 83004 | 270278 | 3602955 | 1565495165 |
| ALLPATHS-LG | 9349 | 8778 | 5252889 | 20294336 | 1053394660 |
| SOAPdenovo | 6389428 | 61403 | 21972 | 3531297 | 2225962440 |
| **Rescaffolder** | | | | | |
| BESST | 6636 | - | 5483466 | - | 1047904772 |
| REAPR | 18512 | - | 225080 | 1461397 | 1050885999 |

* NG50 = length of the longest scaffold such that the sum of all the scaffolds longer than it is at least 50% of the estimated genome length.

Table 2. BUSCO searched for 429 conserved eukaryotic genes in each assembly.

| Assembler | Complete | Duplicates | Fragmented | Missing |
|---|---|---|---|---|
| **abyss** | 206 | 8 | 59 | 164 |
| **ALLPATHS-LG** | 277 | 5 | 26 | 126 |
| **soapdenovo** | 40 | 1 | 45 | 344 |

Table 3. ALLPATHS-LG library coverage report

| Insert size | # reads | % used | Sequence coverage* | Physical coverage** |
|---|---|---|---|---|
| 180 | 747,806,858 | 87.5 | 80.5 | 59.1 |
| 3000 | 25,533,486 | 55.1 | 1.5 | 20.7 |
| 6000 | 97,119,618 | 65.1 | 5.8 | 184.9 |

* Average number of times a base is read

** Average number of times a base is read or spanned by a mate-pair library

Table 4. Mate-pair library summary. Raw read count, expected versus true insert sizes, quality filtering results, and the percentage of reads mapped to the *M. flava tschutschensis* reference genome.

| Exp. insert size (bp) | Mean (stdev) | # raw reads | % reads MP + quality filtered | % reads unknown + quality filtered | % raw reads mapped |
|---|---|---|---|---|---|
| 3000 | 3101 (370) | 18883982 | 40.2% | 25.8% | 92.5% |
| 6000a | 6813 (574) | 45800932 | 44.3% | 16.7% | 91.0% |
| 6000b | 6507 (552) | 31426451 | 43.9% | 18.5% | 92.5% |

Table 5. *M. flava* repeat content listed by repeat type as identified by RepeatMasker.

| Repeat Type | # elements | Length (bp) | % of sequence |
|---|---|---|---|
| Retroelements | 111730 | 35120047 | 3.34 |
| SINEs | 4202 | 500869 | 0.05 |
| Penelope | 77 | 14465 | 0.00 |
| LINEs: | 80383 | 23417344 | 2.23 |
| CRE/SLACS | 0 | 0 | 0.00 |
| L2/CR1/Rex | 80230 | 23381975 | 3.42 |
| R1/LOA/Jockey | 0 | 0 | 0.00 |
| R2/R4/NeSL | 24 | 12243 | 0.00 |
| RTE/Bov-B | 18 | 1086 | 0.00 |
| L1/CIN4 | 34 | 7575 | 0.00 |
| LTR elements: | 27145 | 11201834 | 1.07 |
| BEL/Pao | 0 | 0 | 0.00 |
| Ty1/Copia | 0 | 0 | 0.00 |
| Gypsy/DIRS1 | 0 | 0 | 0.00 |
| Retroviral | 27052 | 11185613 | 1.06 |
| DNA transposons | 8438 | 1288732 | 0.12 |
| hobo-Activator | 1301 | 217888 | 0.02 |
| Tc1-IS630-Pogo | 286 | 50208 | 0.00 |
| En-Spm | 0 | 0 | 0.00 |
| MuDR-IS905 | 0 | 0 | 0.00 |
| PiggyBac | 0 | 0 | 0.00 |
| Tourist/Harbinger | 2144 | 199585 | 0.02 |
| Other | 0 | 0 | 0.00 |
| Rolling-circles | 0 | 0 | 0.00 |
| Unclassified | 1440 | 254400 | 0.02 |
| Total interspersed repeats | - | 36663179 | 3.49 |
| Small RNA | 946 | 122700 | 0.01 |
| Satellites | 1298 | 189361 | 0.02 |
| Simple repeats | 9596 | 1186099 | 0.11 |
| Low complexity | 303 | 64370 | 0.01 |

Table 6. Summary of *M. flava* annotation.

| | | |
|---|---|---|
| **Number of genes/mRNA** | | 9463 |
| | **# exons** | 75256 |
| | **# introns** | 65793 |
| | **# CDS** | 9463 |
| **Overlapping genes** | | 879 |
| **Total gene/mRNA length** | | 196388885 |
| **Total exon length** | | 12823340 |
| **Total intron length** | | 183697131 |
| **Total CDS length** | | 12787233 |
| **Mean gene/mRNA length** | | 20753 |
| **Mean exon length** | | 170 |
| **Mean intron length** | | 2792 |
| **Mean CDS length** | | 1351 |
| **% of genome covered by genes** | | 18.7 |
| **Mean exons per mRNA** | | 8 |
| **Mean introns per mRNA** | | 7 |

Figure 1. Kmer profile of *M. flava tschutschensis* raw reads. Ideally, this plot should be a gaussian distribution. However, it is normal to see a peak for x = 1 (errors) and a second peak close to the expected coverage. If a genome is highly heterozygous, a third peak will be seen at half the expected coverage.

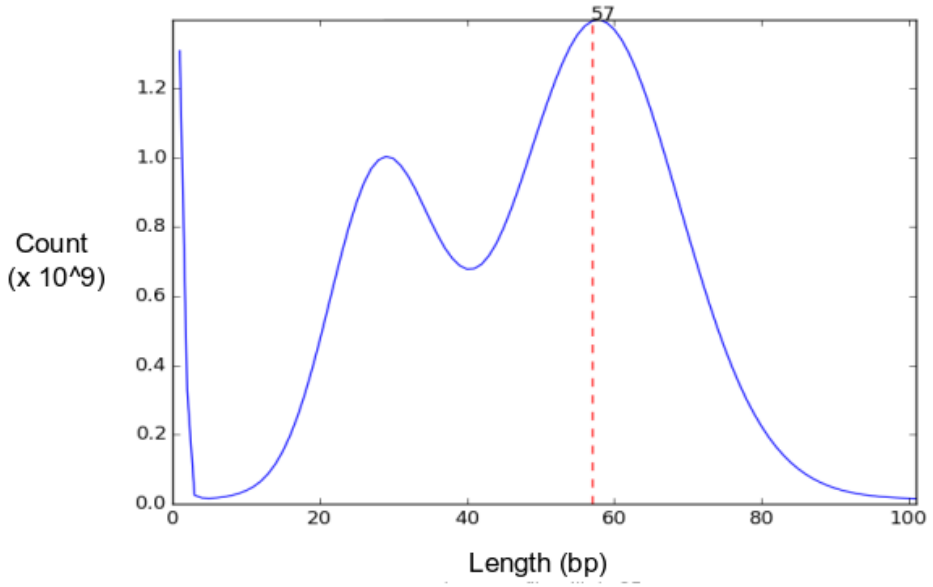


Figure 2. Quality control plots for the genome assembled by abyss.

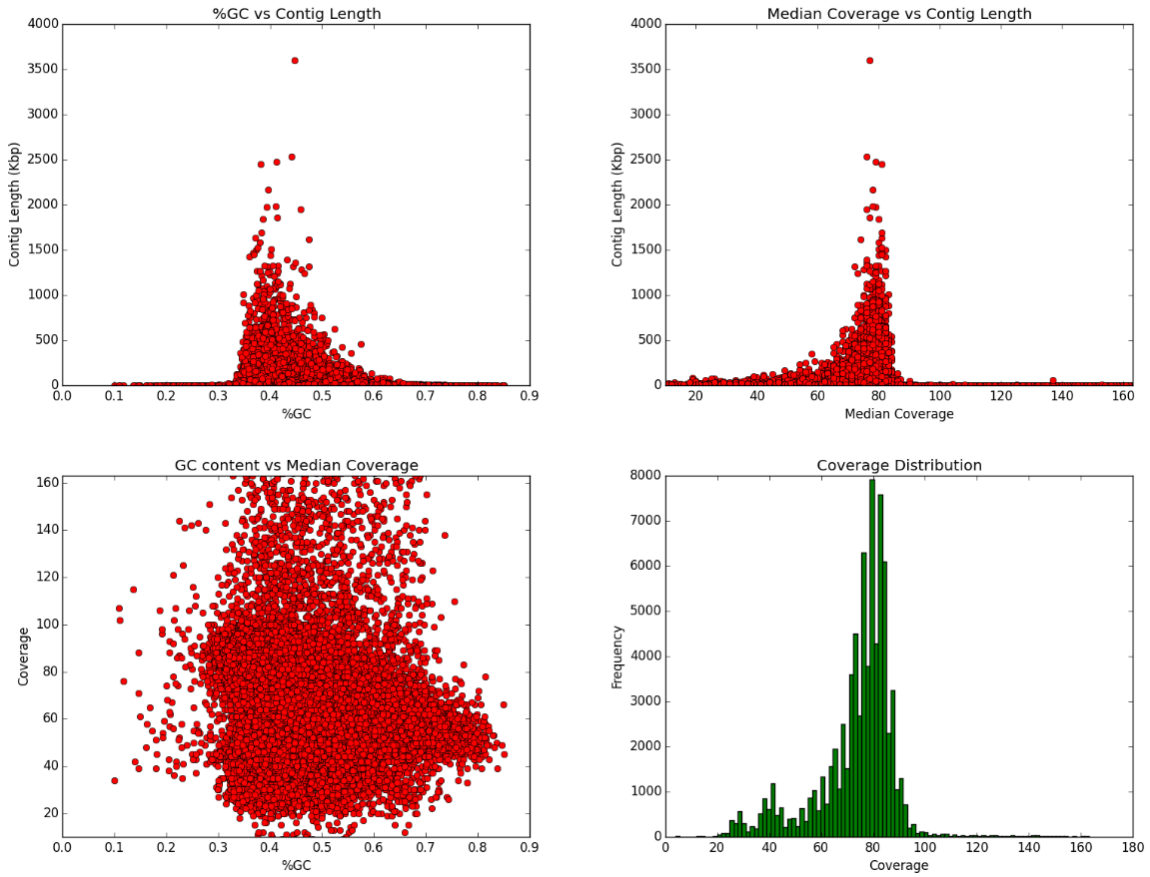


Figure 3. Quality control plots for the genome assembled by ALLPATHS-LG.

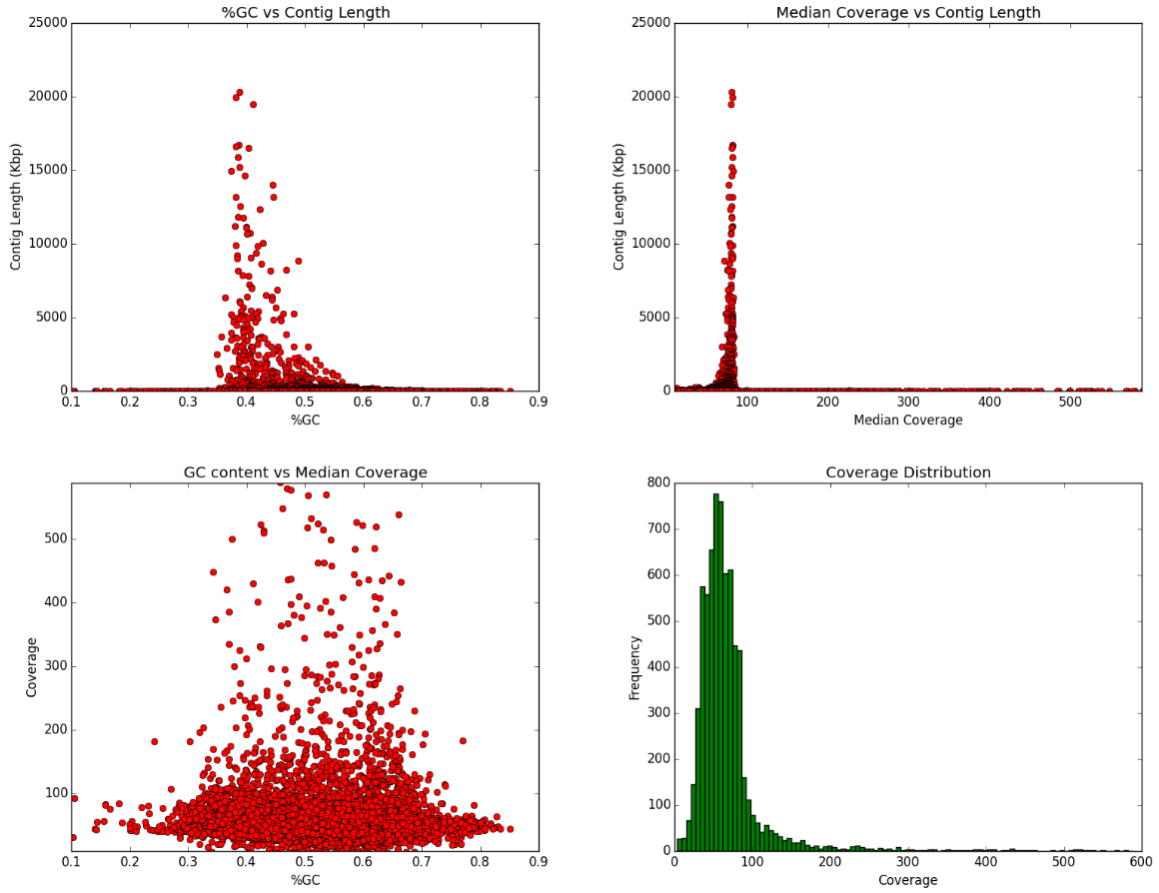


Figure 4. Quality control plots for the genome assembled by soapdenovo2.

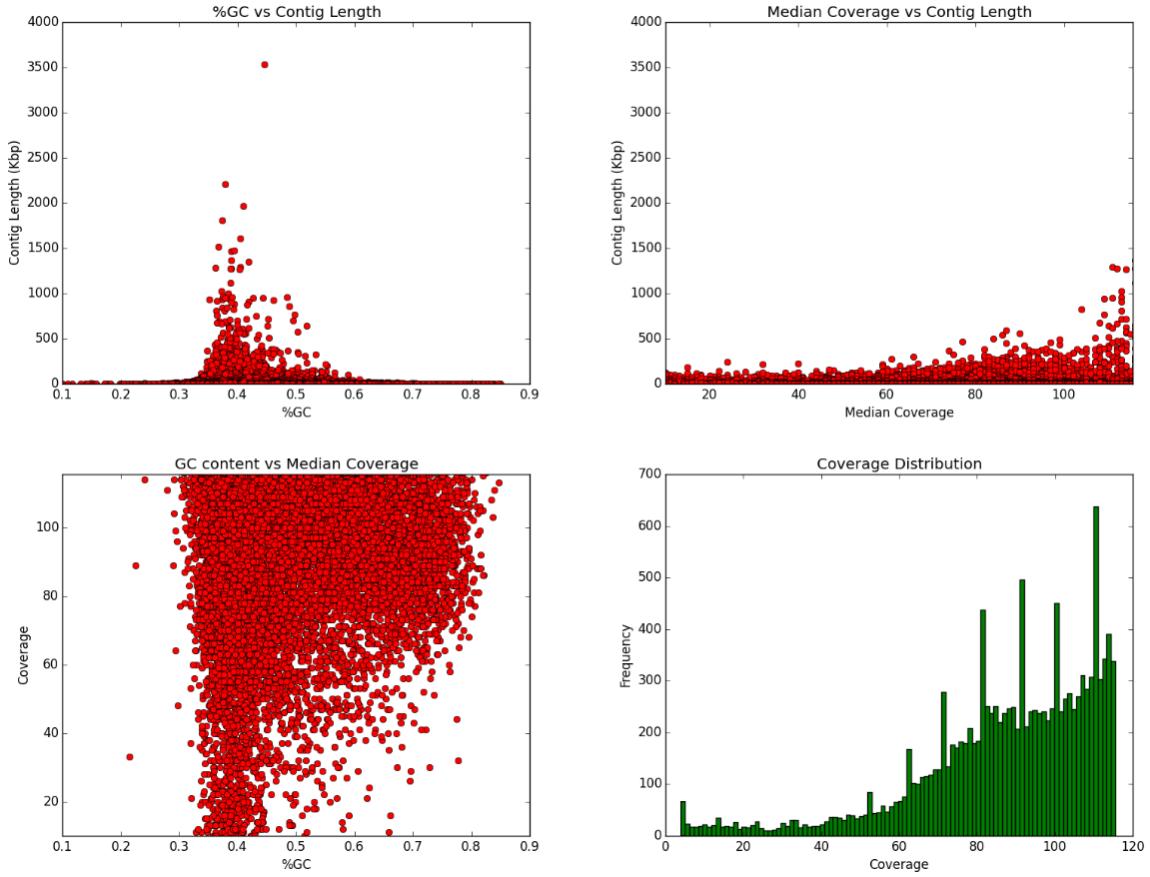


Figure 5. Feature-response curve characterizing the sensitivity (coverage) of the sequence assembler output (contigs) as a function of its discrimination threshold (number of features). An assembler that reconstructs a higher portion of the genome with fewer features is desirable and this translates to a sharper FRCurve.

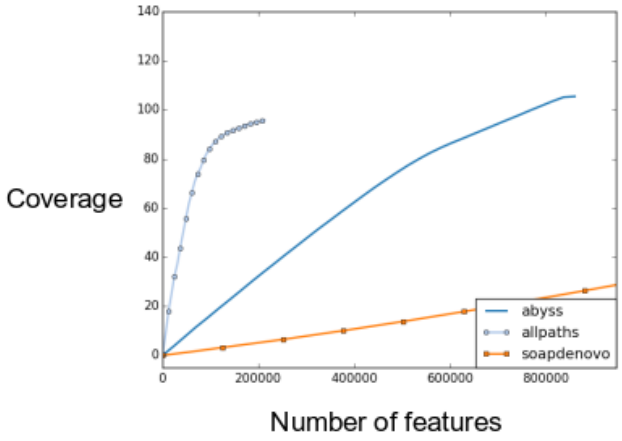


Figure 6. Depiction of *M. flava* scaffolds mapping to zebra finch chromosomes. Colors indicate mapping of forward (black) vs. reverse (red) scaffolds. No coverage is shown as gaps in color (white). The absence of linkage groups 2 and 5 (LG2, LG5) is expected, as they are zebra finch specific.

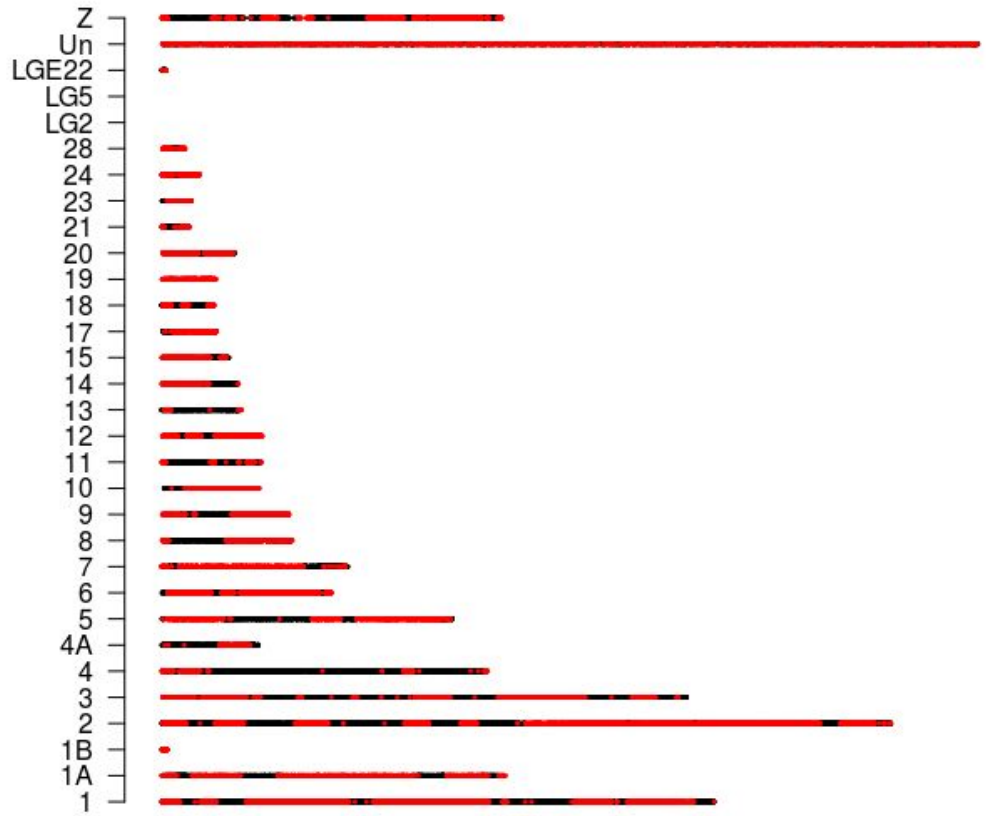
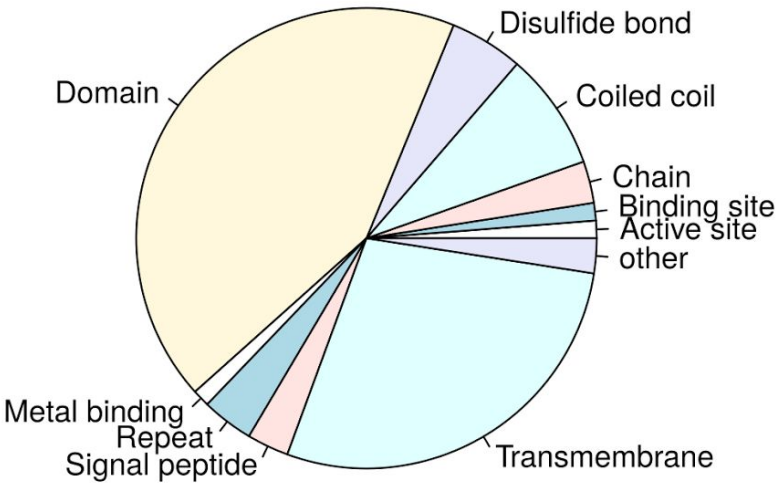


Figure 7. Characterization of proteins by Uniprot classifications. ‘Other’ category consists of all classifications comprising < 1% of the genome (calcium binding, cross-link, DNA binding, initiator methionine, intramembrane, lipidation, nucleotide binding, motif, topological domain, zinc finger, glycosylation, and modified residue).



**Chapter 3:** Quantifying phenotypic and genotypic diversity in the white and yellow wagtail species complexes (Aves: *Motacilla*) using whole-genome resequencing

Rebecca B. Harris

## Introduction

Population divergence is often started with a period of geographic isolation during which there is limited gene flow between populations and the potential for divergent selection. Traditionally, it was expected that populations under divergent selection would experience and fix a different series of mutations at genes responsible for trait differences. If selection occurs on traits important in reproductive isolation, then these populations are more likely to develop reproductive barriers [1]. Many studies have focused on a set of candidate genes thought to have a functional role in diverging traits. However, this candidate gene approach may be ineffective for studying complex traits which involve multiple genes mediated by their interaction with environmental factors. Furthermore, the underlying genetic control of many traits is still unknown. Studies focusing on a specific subset of genes may fail to elucidate differences important in population differentiation. With the advancement of sequencing technologies, studies utilizing genome-wide scans have identified previously unrecognized candidate genes for adaptation and trait differences [2]. Such studies have also provided evidence that the signal of differentiation across the genome is widespread and heterogeneous [3].

Early in the speciation process, patterns of divergence may lack consistent signal, due to the complex interactions of drift, selection, recombination, mutation, introgression, and incomplete lineage sorting. Genome scans provide a high-resolution picture of genome-wide divergence and are not biased towards functional regions. Regions of high genomic differentiation can arise by means unrelated to reproductive isolation or speciation [4]. To understand whether a trait is under selection, one must first understand the evolutionary processes shaping the genome. For example, demographic history can produce patterns similar to that of selection and can lead to erroneous conclusions about the process of evolution [5]. In addition, genomic architecture may complicate the detection of genes underlying population differentiation [6]. Recent studies have begun to parse population specific divergence from shared evolutionary history by sampling several populations at various stages of

population divergence and combining analyses of loci under divergent selection with genome-wide scans and robust estimates of demographic history.

By coupling candidate gene approaches with characterization of patterns of genetic variation across the genome, there is far greater power to reveal evolutionary processes driving the non-random accumulation of genetic differences between diverging populations [7]. Genome scans allow us to detect previously unknown genomic regions that associate with traits of interest.

*Genes of interest*

In songbirds, if selection occurs on traits important in reproduction, such as plumage, song, or habitat choice, these populations are more likely to develop reproductive isolation [8]. There are many examples of natural and sexual selection driving geographic variation in plumage. Sexually selected plumage traits are important for species recognition and have the potential to rapidly diverge between allopatric, ecologically equivalent populations [9]. Because mate preferences in birds are often learned through imprinting on parental phenotypes, barriers to gene flow can arise without genetic coupling between mating trait and preference [10]. The lack of genetic structure concordant with phenotype suggests that this may be the case in the white (*M. alba*) and yellow (*M. flava*) wagtail species complexes.

To elucidate the genetic basis of plumage in wagtails, we targeted genes known to contribute in some way to pigmentation pathways. The underlying genetic variants associated with pigmentation phenotypes are relatively well characterized. Melanin-based coloration has repeatedly been shown to be heritable and associated with few mutations in melanogenic genes [11–13]. Knowledge about the molecular basis of carotenoids deposition is still limited, although carotenoid-based pigments tend to be more phenotypically plastic and can be environmentally dependent [14].

Another trait that may be important in wagtail evolution is migration. Migration is a labile trait [15] and the evolution of new migratory behaviors has the potential to facilitate ecological, morphological, physiological and life-history evolution [16]. In particular, migratory direction and the timing of return to the breeding ground may have large fitness consequences. Arrival date and a number of other migratory behaviors are associated with genetic differentiation [17,18]. A few studies have demonstrated a link between sexually selected traits and migratory behavior [19,20]. Both the white and yellow wagtail species complexes are highly migratory and distributed across a large south-north expanse during the breeding season. If migration is playing an role in population differentiation, then it may be revealed by searching for fixed differences in known migration genes. Circadian rhythms help birds anticipate upcoming changes in environmental conditions [21] and act as timekeepers for avian migration. The molecular basis for circadian rhythms is highly conserved [22]. We target genes that may be involved in the presumed migration differences within the wagtail species complexes.

*Demographic history*

Characterizing genomic differentiation requires an understanding of demographic history. Non-adaptive demographic processes will have a universal and random effect on the genome [23], resulting in genetic structure and differentiation unrelated to phenotypic divergence. Isolation followed by founder events and population expansion can leave distinct signatures on the genome that, unaccounted for, may resemble divergent selection [24]. Given dynamic climatic change during the Pleistocene, it is necessary to account for demographic history to understand population divergence in wagtails. The Quaternary (2.4 mya to present) has been characterized by large climatic oscillations with warm interglacials and cold glacial periods [25]. This climatic cycling is thought to have caused repeated population expansions and contractions.

Due to missing data and a sparse data matrix, we were unable to optimize demographic models reflecting the population fluctuations characteristic of many species impacted by the Pleistocene (Chapter 1). Here, we take advantage of the whole-genome resequencing data generated for outlier detection and apply it to the relatively new Pairwise Sequentially Markovian Coalescent (PSMC) model [26]. PSMC has been successfully used to show population fluctuations in a number of studies [27,28] and therefore should provide a clearer picture of past demographic events affecting the white and yellow wagtail species complexes. PSMC assumes that changes in the TMRCA is due to recombination, where the probability of recombination is dependent on the mutation rate, recombination rate, and the relative effective population time at a given interval. From this information, PSMC estimates effective population size ($N_e$) across all past coalescent intervals. An advantage of the PSMC model is that it uses one genome from a single individual to make inferences. Comparing population trajectories among individuals of the same and different subspecies can reveal something about population history. For examples, if trajectories of individuals from the same subspecies are congruent, but different across subspecies, then it suggests that subspecies designations reflect biologically meaningful lineages [27,29].

By sequencing multiple genomes from across the white and yellow wagtail species complexes, we have the opportunity to characterize genomic differentiation, identify important genomic regions that may be contributing to population differentiation, and get a more robust estimate of their demographic history. The goal of this study is to provide a platform for future genomic studies seeking to understand how demographic processes and selection have shaped phenotypic diversity in wagtails.

## Methods

### *Sampling and sequencing*

We chose five subspecies from each species complex (Table 1), including representatives from the subspecies groups (hereafter referred to as populations) estimated in Chapter 1. (Note: we found three yellow wagtail populations (west, northeast, southeast) in Chapter 1. Due to sampling restrictions, we were only able to include one subspecies representative from the northeastern and southeastern populations. Therefore, in the remainder of the manuscript, we collectively refer to these as the "eastern" population.)

To guide our choice in sample selection, we characterized plumage diversity within the two species complexes. We used a data matrix of 30 discrete plumage characteristics generated by Alström *et al.* (2002) to calculate a distance matrix using the R package claddis [30]]. Prior to principal component analysis (PCA), we rescaled raw distances against the maximum possible observable distance, as described in Lloyd *et al*. (2016). We attempted to include representatives from subspecies with a) similar plumage but disjunct distributions, or b) different combinations of discrete plumage characteristics that would allow exploration of genome-wide associations. However, we were limited by the availability of high quality tissue samples with associated specimen skins. From each subspecies, we selected 1-3 individuals per subspecies from the Burke Museum collections. Individuals were assigned to subspecies according to the criteria outlined in Alström *et al.* [31]. For plumage analysis and to ensure equal coverage of both autosomes and sex chromosomes, only males were used.

DNA was extracted from tissue using standard phenol-chloroform extractions, quantified with the broad-range Qubit fluorometer (Life Technologies) protocol, and quality checked for length and fragmentation with a TapeStation (Agilent). DNA library construction and sequencing was completed into two separate batches by the Brigham Young University Sequencing Center and UC Berkeley's QB3. All individuals were prepared with Illumina's TruSeq kit with short-insert (350 bp) libraries. Three lanes of sequencing was performed using 125 bp paired-end reads on the Illumina HiSeq2500 platform.

*Alignment and filtering*

We filtered raw reads using a combination of CutAdapt [32], Trimmomatic [33], and a custom C script (T. Flouri). The quality of reads was assessed in FastQC. Quality filtered reads were aligned to the *M. flava tschutschensis* reference genome (see Chapter 2) with bowtie2 [34], specifying insert sizes between 200 and 500 (using *-I* and *-X*, respectively). After converting to sorted BAM format for efficient storage and assigning read group names, duplicates were marked with picard tools. To improve variant calls, we merged BAM files from the same species before conducting indel realignment with GATK [35]. Variant calling was performed with samtools mpileup [36] and bcftools. In both vcf and fastq files, only sites with quality greater than 20 (*-Q*) were used. Further quality filtering was conducting using a combination of vcftools [37] and SnpSift [38]. Depending on the analyses, the depth of coverage and missing data threshold varied.

Low coverage regions risk being called homozygous because the other allele may not have been sequenced by chance. High coverage sites are potentially the result of collapsed regions in the assembly. Therefore, we filtered out sites with less than 5x coverage or more than 100x coverage when constructing consensus fastq/fasta files with the samtools vcfutils function.

*Population structure*

To explore population differentiation, we used a complete data matrix (0% missing data) with a minimum 10x depth of coverage and excluded SNPs within 3 bp of indel regions. We calculated hierarchical genome-wide Fst using ANGSD [39]. Fst was calculated separately for each locus and weighted pairwise Fst was calculated between subspecies.

To explore whether denser genome-wide sampling provides additional information that alters estimates of population structure or relationships, we repeated some of the analyses conducted in

Chapter 1 with our resequencing data. Population structure was estimated with DAPC in adegenet [40]. Subspecies relationships were estimated from a concatenated gene tree in RAxML using the Lewis correction with a HKY85 model and 1000 bootstrap iterations [41]. So as not to violate the assumption of unlinked sites in both analyses, we excluded SNPs within 10 kbp of each other. To account for the potential variance in subsetting the full data matrix, we conducted both analyses with five independent draws of SNPs separated by 10 kbp.

*Demographic inference with PSMC*

Estimates of effective population size are highly sensitive to the type of genomic data used and the depth of genomic coverage. Since the effective population size of the Z chromosome is only ¾ of that of autosomes, Z linked genes will display reduced effective population. Therefore, we removed all Z linked scaffolds from the PSMC analysis. These were identified using the NUCMER alignment of scaffolds to the zebra finch genome (Chapter 2). A recent empirical study on flycatchers found that below 10x coverage PSMC failed to reconstruct the demographic history found at higher coverage [29]. Others studies have suggested a threshold as low as 5x [27]. To determine the threshold in the present study, we conducted PSMC using a 5x, 10x, and 20x threshold for *M. alba lugens* (alb57). If demographic analyses was sensitive to coverage, then we expected a change in the shape or position of the PSMC curve. However, we did not see a biologically meaningful change in our PSMC results and therefore use a 5x threshold for all other analyses. The *M. f. lutea* individual was excluded as genome coverage was too low for reliable PSMC analysis.

The first step of the PSMC analysis (fq2psmcfa) is to split the sequence into 100 bp bins, where a bin is considered heterozygous if > 10 bp were called and at least one base is heterozygous. We then used this new file to infer demographic history using 30 iterations (-*N*). $N_e$ was inferred across 34 free atomic time intervals (4+30*2+4+6+10), meaning the first population size-parameter spans the first four atomic intervals, each of the next 30 parameters spans two intervals, while the last

three parameters spans four size and ten intervals (*-p*). These settings have been used to analyze other passerine species with similar divergence histories [29,42].

To explore how parameter choice influenced demographic estimates, we chose a range of values for the upper limit of the TMRCA (*-t* = 5, 15) and the mutation/recombination ratio (*-r* = 1, 3, 5). Our results were not sensitive to these parameter choices, so our final analyses were conducted using a TMRCA of 5 and a ratio of 3. To estimate variance in Ne, we performed 100 bootstrap replicates by randomly sampling with replacement 5 Mb sequence segments. A generation time of 1 year and a mutation rate of 2.3x10^-9 [43] was used to convert PSMC results into real-time units.

*Plumage & migration*

We used two methods to explore the presence of fixed differences in genome sequene. First, we explored the prevalence and pattern of SNPs within genes of interest. Second, we conducted a sliding window analysis of fixed differences and Fst.

*Candidate genes* - Using the Gene Ontological database, we generated a list of 84 genes (Table 2) implicated in migration (circadian rhythm) [44] and the plumage pigment (melanin and carotenoid) pathways [12,14]. Each gene name was searched against the reference genome annotations (Chapter 2). Only 49 matches were found. Missing genes were likely due to a fragmented and incomplete reference genome (discussed in Chapter 2). This was substantiated by our inability to align more than ~50% of our WGS data to the reference (Table 1). For each match, we extracted positional information which we then used to subset consensus fastq files. Because individual-specific indels are carried over to fastq alignments, we checked and refined the alignment of single genes visually in JalView [45]. For genes exhibiting a large number of SNPs, we calculated the probability of seeing a cluster of SNPs by calculating the average interval between SNPs across all scaffolds. Then we conducted a t-test to determine whether the interval seen on gene of interest was significantly different than average.

When looking for candidate genes, it is important to include enough samples from each group to ensure that what appears to be a fixed difference is not actually an intra-group polymorphism. The minimum number of samples needed to confidently call a fixed difference depends on the degree of intra-group diversity. Normally this is encapsulated in estimates of theta, but our current sampling was too small to reliably estimate theta for subspecies. Therefore, we calculated the percent genome-wide heterozygosity. To do this, we implemented the heterozygosity estimator in ANGSD [39] with a 20x depth of coverage threshold (20x) so as to minimize the effects of sequencing error and ensure sampling of both alleles.

*Sliding window* - Outlier tests are a common way to detect regions of the genome that have a higher Fst than expected under genetic drift alone. Such tests have the potential to identify regions under divergent selection (high Fst) or balancing selection (low Fst). However, non selective forces, such as dispersal, population structure, or demographic history, can impact genome-wide Fst values [24].

A recent simulation study demonstrated that a large number of SNP markers (100-1000) can compensate for small sample (individuals) sizes (n = 2-6) [46]. However, this study used a simple model of population history and may not reflect the complexity of real demographic scenarios. Assuming an avian mutation rate of 2.3 x 10^-9 [43] and our estimated population divergence for the split between the "eastern" and western yellow wagtail populations, we calculated the size of a sliding window that would on-average provide enough SNPs. Using a divergence time of 150 kya (Chapter 1 & results), we show that a 300 kbp sliding window is needed to expect an average of 103.5 SNPs per window. That the N50 of our reference genome is smaller than the required sliding window (225 kbp; Chapter 2) and the resequencing data only covered about 50% of the reference genome makes it unlikely that Fst sliding windows between subspecies or white wagtail populations (with more recent divergence times, see Chapter 1) would be reliable. Therefore, we limit the Fst sliding window analysis to just between the "eastern" and western *M. flava* populations.

Instead, we performed a sliding window analysis of fixed differences between subspecies using custom R scripts. We used the coordinates provided by the NUCMER reference genome alignment to the zebra finch genome (Chapter 2) and the frequency of fixed SNPs calculated across each 10 kbp interval. We did not collapse data from scaffolds > 10 kbp. We conducted all pairwise comparisons between subspecies and populations. For each comparison, sites with missing data were removed.

**Results**

*Summary statistics*

As expected, the raw number of sequencing reads varied across individuals and sequencing lanes (Table 1). Across individuals and species, the indel rate was 0.001-0.002, the mean mismatch rate was 0.022, and only half of our reads were mapped to the reference genome. Compared to similar studies, this mapping rate is low (per. comm. Alexander Suh) and likely indicative of issues with the reference genome, such as a high degree of repetitive sequence.

*Population differentiation*

Taken together, these analyses suggest that the genome-wide signal of population structure is more pronounced in the yellow wagtail species complex. Genome-wide Fst values are higher between *M. flava* subspecies than between *M. alba* subspecies (Table 3).

*M. flava* - Across all five datasets, both DAPC and RAxML converged on the same results. Our PCA results mirror the east-west divide in *M. flava* recovered in Chapter 1 (Fig. 2). The same pattern was observed for *de novo* population structure estimation. The PCA did not support a further splitting of the "eastern" population into northern and southern groups, likely due to our limited

sampling. RAxML corroborates these findings. RAxML found high support for the monophyly of *taivana* and *tschutschensis* and their sister relationship (Fig. 3). The western subspecies are monophyletic (*flava, feldegg, lutea*), nor were any of the subspecies within it. The lack of monophyly of the western group is solely due to the placement of *M. f. lutea.* However, this sample had the lowest coverage of any (Table 1) and missing alleles may be impacting the analysis.

*M. alba* - Neither DAPC nor RAxML recovers the same east-west *M. alba* split seen in Chapter 1. Of the sampled subspecies, only *M. a. lugens* was found to be monophyletic.

*Demographic analysis*

We were able to reconstruct effective population size spanning nearly all of the Pleistocene period (10 kya - 2 mya) (Fig. 4-6). We depict bootstrapped PMSC results according to subspecies (Fig. 4 & 5) to demonstrate whether there was a subspecies-specific population trajectory. To compare across species complexes, we depict only one individual (highest coverage) per subspecies with bootstrap intervals in Figure 6.

*M. alba* - PSMC found largely concordant demographic estimates within subspecies (Fig. 5). Across subspecies, population trajectories are very similar. Around 200 kya, all *M. alba* subspecies underwent population expansion. Populations continued to increase through the start of the last ice age (110 kya) and reached their maximum at 60 kya and 85 kya in the western and eastern populations, respectively. The eastern population reached a $N_e$ peak of approximately 2 million, whereas the western population peaked at approximately 3 million breeding individuals. Population size decreased until the most recent estimate 10 kya in all but *M. alba.*

*M. flava* - In contrast to the uniformity of the *M. alba* subspecies, the *M. flava* subspecies show trajectories that are consistent within but distinct subspecies. Across all subspecies, population growth begins around 1 mya but reaches its peak at variable times. First, we discuss demographic trends in the northeastern and southeastern subspecies, *tschutschensis* and *taivana*, respectively.

Between 15-100 kya there is some uncertainty in *taivana* $N_e$ - in general, $N_e$ appears to steadily increase to ~2-2.5 million, followed by an abrupt 100 fold decrease in Ne. In contrast, *tschutschensis* $N_e$ peaks (~2.5 million) at 125 kya and then steadily declines, reaching a quarter of its size by 16 kya.

The western subspecies (*beema*, *flava*, and *feldegg*) show striking differences in population trajectory. The classification of *beema* as a subspecies separate from *flava* has been questioned [31]. In Figure 4, we show the *beema* separate from *flava*, as the PSMC analysis infers such different demographic histories. Notably, *flava* appears to have increased in $N_e$ until recent times, whereas *beema* reached a peak ~110 kya, experienced a slight decline, followed by suddenly doubling $N_e$ ~32 kya and then a 6-fold decrease ~20 kya. The two *feldegg* individuals show a similar population curve with an steady increase to an $N_e$ maximum, followed by a bottleneck and then recovery. However, the time estimates for *feldegg* do not align - likely due to low coverage in one of the samples.

Overall PSMC estimates population sizes in *M. alba* are much smaller than those of *M. flava.* Across all subspecies *M. alba* had half the rate of heterozygosity of *M. flava* (Table 1). This is in accordance with the neutral theory, which states that population with larger $N_e$ should have increased heterozygosity [47]. The heterozygosity rates found in wagtails is of the same order of magnitude as other bird species [48].

*Genomic differentiation*

The percentage of genome-wide fixed differences in subspecies varies (Table 4). As a point of reference, humans and chimps diverged from their common ancestors 5-7 million years ago and their genome consists of ~1% fixed differences [49]. The genome wide pattern of fixed SNPs also suggests a higher degree of divergence in *M. flava* compared to *M. alba*.

*Genes of interest*

*M. flava* - Few genes stand out as highly differentiated. Of those, *TYR* differentiates *taivana* + *tschutschensis* from the other yellow wagtails.The *TYR* transcript (intron + exon) contains 121 fixed SNPs (mean interval = 121 bp) holding this pattern. The probability of finding a cluster of SNPs distinguishing *taivana* + *tschutscensis* from the other wagtail populations is extremely low (p-value $2.2 \times 10^{-16}$). 18 of these fixed SNPs occur in the coding region of *TYR*, making up 10 changes in amino acid residues.

Another gene, one involved in migration, that finds the same pattern at a highly unlikely cluster is *Nfil3*. This gene has 54 fixed SNPS within its transcript, seven of which are within the coding region.

*M. alba* - We found no such high density clusters of SNPs in *M. alba* in neither the pigment or migration related genes. We find no shared SNPs in any of the *M. alba* migration genes. We find one fixed SNP in *TYR* and *MC4R* distinguishing *personata* + *alba* and *alba*, respectively.

*Sliding window analysis*

*M. alba* - The lack of differentiation between the eastern and western *M. alba* populations is striking (Fig. 7). The pairwise subspecies analyses showed a greater number of divergent regions. Even so, the maximum value number of fixed differences in any 10 kbp window was 12. There is no apparent outlier in the pairwise analyses of eastern vs. western subspecies. Some chromosomes showed a higher degree of differentiation (notably, chromosomes 1, 1A, 2, 5, Z) but between these there was no clear outlier chromosome.

*M. flava* - The *M. flava* sliding window analyses revealed high differentiation on the avian sex chromosome (Z) compared to the autosomal chromosomes (Fig. 8) for both “eastern” vs. western and the northeastern vs. southeastern comparison. This pattern is absent from the pairwise comparisons between western subspecies.

We found heterogeneity in Fst across the "eastern" and western *M. flava* populations (Fig. 9). The pattern of differentiation reveals a similar pattern to the analysis of SNPs (Fig. 8). However, the Fst analysis found differentiation peaks on chromosome 10 and 18 undetected by the SNPs sliding window. This could be to a number of differences between the analyses: window size (300 kbp vs. 10 kbp), filtering method, and tolerance for missing data.

**Discussion**

We documented substantial genomic heterogeneity in sequence divergence in the yellow wagtail species complex, with fixed SNPs located throughout the genome and some occurring in large clusters. In contrast, we found low levels of divergence in the white wagtail species complex with no clusters of fixed differences. This agrees with the timing of divergence between the species; we find a strong signal for "eastern" and western populations. We also find some evidence for subspecies specific trajectories in our demographic analyses. The same cannot be said for the white wagtails. We do not recover strong evidence for population structure consistent with the findings of Chapter 1 and there is no genomic pattern consistent with subspecies specific trajectories.

*Demographic history*

Overall, we confirm that PSMC is a useful tool at pinpointing the time of lineage splitting at the population level. Convergent PSMC curves of all five *M. flava* subspecies up until ~100 kya can be interpreted as shared ancestry and demographic history. After this time, differing population trajectories demonstrate that the five *M. flava* subspecies were experience distinct population trajectories. That populations split around 150 kya is consistent with our time-calibrated species tree and *dadi* estimates in Chapter 1. In contrast, the PSMC results of *M. alba* are more subtle and distinct population trajectories are not immediately apparent. In combination with our Chapter 1 population

structure results, we can deduce that the eastern and western populations experienced different trajectories. This is apparent from a slight shift in timing of $N_e$ peak and a smaller $N_e$ in the eastern subspecies. The shared trajectory until approximately 100 kya suggests *M. alba* lineage splitting at this time.

The timing of the last glacial maxima (LGM; ~20 kya [50]) coincides with the dramatic $N_e$ decline in both *M. f. beema* and *M. f. taivana*, and the $N_e$ low point seen across all *M. alba* subspecies and *M. f. tschutschensis*. During the last glacial period (110-40 kya) leading up to the LGM, all *M. alba* experienced a population decline likely due to loss of suitable habitat. However, the three western *M. flava* subspecies increased during this time suggesting that they were unaffected by habitat loss. One region of the current *M. f. flava* range known to be covered in an ice sheet from 90 kya to the LGM was Scandinavia [51]. The final increase in $N_e$ starting at this point may indicate population expansion and colonization of this area. Future studies should prioritize the inclusion of the most southern subspecies from the *M. alba* species complex.

Combined with *dadi* analyses (Chapter 1), we appear to have compelling evidence that the Eurasian *Motacilla* were not affected by cyclical climate change. Other studies have used PSMC to document fluctuations in population size concordant with historic climate oscillations [27]. However, both of these methods may fail to accurately portray historical population fluctuations if there was a severe population bottleneck and a drastic reduction in genetic diversity. Both methods rely on retention of some genetic variability, but if the bottleneck was severe and nearly all variability was lost, then we might expect only one episode of population expansion and contraction. Future studies on wagtails should utilize simulations to demonstrate how severe such a bottleneck would need to be and whether there has been sufficient time to allow the recovery we see today.

*Genes of interest*

We identified two genes that may be playing a role in driving population differentiation in the yellow wagtail species complex. While we can conclude that divergence at these loci occurred before the completion of speciation, it is unknown whether these genes are contributing to the evolution of reproductive isolation or their effect size - therefore, their role as speciation genes still needs to be established [52].

*Nfil3* plays a role in the regulation of circadian rhythm gene expression [53] and contains a high density of SNPs distinguishing the eastern and western *M. flava* populations. With such a vast range, it is possible that *M. flava* has a migratory divide - with eastern and western populations differing in either the timing or direction of migration. Migratory divides have been demonstrated in a number of other species [54–57]. To further explore whether there is a role for migration in driving population differentiation, cage orientation, gene association experiments, and studies of birds on the wintering ground should be done [17].

The *TYR* gene belongs to the tyrosinase-related protein gene family along with *TYRP1* and *TYRP2*. They have known involvement in pigmentation, specifically the regulation and control of melanin synthesis, and share the same signalling sequence, including two binding sites responsible for their catalytic activities [58]. These genes have the potential for functional polymorphisms and could explain natural variation in pigmentation phenotypes [59] The expression level of these genes has also been shown to play a major role in determining plumage color in other birds [60].

White wagtails show differentiation in color pattern, not color itself, and this may be due to local differentiation in gene expression rather than amino acid changes [61]. For example, differences in expression distinguish crows with grey vs. black backs [62]. Hormone levels can also play a role in the intensity of pigmentation and pattern [63]. However, single gene mutations of large effect have been shown to cause plumage differences [11] and we were unable to retrieve all the genes important in the melanogenesis pathway (most importantly, *Agouti [64]*). Yellow wagtails display color pattern patches (darker yellow, black, and white) that are likely a result of melanin deposition as well.

While wagtails may benefit from a more complete candidate gene panel, it is also likely that this group would also benefit from future transcriptomics work. With such a small sample size, we cannot reliably interpret our results to conclude that these genes are under selection, as they may instead reflect demographic processes or population structure. Increased sampling at the individual level will allow for calculation of Fst (a more appropriate way of accounting for differentiation than number of fixed SNPs), tests of neutrality, recombination, and linkage disequilibrium. Negative values of Tajima's D or Fu's Fs may help elucidate whether the patterns seen are due to demographic processes rather than selection.

*Sliding windows*

Our sliding window analysis reveals high differentiation on the avian sex chromosome (Z) in *M. flava* but not *M. alba*. Until recently, it was thought that mutations on the sex chromosome were more likely to be fixed because of sexually antagonistic fitness effects and more exposure to selection. Both theory and empirical studies have shown that genes on the sex chromosomes play an important role in reproductive isolation. If the Z chromosome harbors genes involved in female preference and flashy male traits, then this should promote the rapid evolution of sexually selected traits [65]. However, the Z chromosome has a lower effective population size compared to autosomes and this can lead to an increased rate of functional change due to a reduced efficacy of purify selection [66]. Because of this, genetic drift would then have a greater potential to fix mildly deleterious alleles. New evidence suggests that drift is primarily responsible for the higher rates of evolution on the Z chromosome of Galloanserae, a group of birds that are separated by nearly 90 million years [67]. Given that *M. alba* divergence is very recent, we would not expect to see divergence on the Z chromosome even if selection or drift was ongoing.

Increased sampling will much improve our ability to conduct Fst outlier scans. Not only will this allow us to use more sophisticated methods which incorporate demographic models into their

estimation, but it will allow us to decrease the size of our sliding window which far exceeded the effects of linkage.

**Conclusion**

Despite their plumage differences known to be involved in sexual selection, both wagtail species complexes are differentiated in very small portions of their genome. These regions are distributed throughout the genome, but do not include candidate genes involved in the pigmentation or migration pathways. Our study provides a platform for future studies seeking to identify the genetic basis of phenotypic differences in wagtails. The genotype-phenotype discord observed in wagtails has also been observed in other systems and has been attributed to recent divergence [64,68] and ongoing gene flow [69]. Wagtail research will benefit greatly from a greater understanding of their natural history and plumage diversity. Association studies will require a more robust understanding of plumage variation both within and across subspecies - future research should prioritize quantification of plumage pattern.

**References**


1. Qvarnström A, Bailey RI. Speciation through evolution of sex-linked genes. Heredity . 2009;102: 4–15.

2. Wray GA. Genomics and the Evolution of Phenotypic Traits. Annu Rev Ecol Evol Syst. 2013;44: 51–72.

3. Parchman TL, Gompert Z, Braun MJ, Brumfield RT, McDonald DB, Uy JAC, et al. The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. Mol Ecol. 2013;22: 3304–3317.

4. Noor MAF, Bennett SM. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity . 2009;103: 439–444.

5. Roesti M, Kueng B, Moser D, Berner D. The genomics of ecological vicariance in threespine stickleback fish. Nat Commun. 2015;6: 8767.

6. Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, et al. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat Commun. 2016;7: 13195.

7. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and the origin of species. Nat Rev Genet. 2014;15: 176–192.

8. Price T. Speciation in Birds. Roberts & Company; 2008.

9. Price T. Sexual selection and natural selection in bird speciation. Philos Trans R Soc Lond B Biol Sci. 1998;353: 251–260.

10. Poelstra JW, Ellegren H, Wolf JBW. An extensive candidate gene approach to speciation: diversity, divergence and linkage disequilibrium in candidate pigmentation genes across the

European crow hybrid zone. Heredity . 2013;111: 467–473.

11. Mundy NI. A window on the genetics of evolution: MC1R and plumage colouration in birds. Proc Biol Sci. 2005;272: 1633–1640.

12. Roulin A, Ducrest A-L. Genetics of colouration in birds. Semin Cell Dev Biol. 2013;24: 594–608.

13. Hoekstra HE. Genetics, development and evolution of adaptive pigmentation in vertebrates. Heredity . 2006;97: 222–234.

14. Toews DPL, Hofmeister NR, Taylor SA. The Evolution and Genetics of Carotenoid Processing in Animals. Trends Genet. 2017;33: 171–182.

15. Helbig AJ. Evolution of Bird Migration: A Phylogenetic and Biogeographic Perspective. Avian Migration. 2003. pp. 3–20.

16. Zink RM. The evolution of avian migration. Biol J Linn Soc Lond. 2011;104: 237–250.

17. Delmore KE, Toews DPL, Germain RR, Owens GL, Irwin DE. The Genetics of Seasonal Migration and Plumage Color. Curr Biol. 2016;26: 2167–2173.

18. Lundberg M, Boss J, Canbäck B, Liedvogel M, Larson KW, Grahn M, et al. Characterisation of a transcriptome to find sequence differences between two differentially migrating subspecies of the willow warbler Phylloscopus trochilus. BMC Genomics. 2013;14: 330.

19. Pape Moller A. Heritability of arrival date in a migratory bird. Proceedings of the Royal Society B: Biological Sciences. 2001;268: 203–206.

20. Friedman NR, Hofmann CM, Kondo B, Omland KE. Correlated evolution of migration and sexual dichromatism in the New World orioles (icterus). Evolution. 2009;63: 3269–3274.

21. Daan S, Aschoff J. Circadian Contributions to Survival. Proceedings in Life Sciences. 1982. pp. 305–321.

22. Cassone VM. Avian circadian organization: a chorus of clocks. Front Neuroendocrinol. 2014;35: 76–88.

23. Knowles LL, Richards CL. Importance of genetic drift during Pleistocene divergence as revealed by analyses of genomic variation. Mol Ecol. 2005;14: 4023–4032.

24. Lotterhos KE, Whitlock MC. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Mol Ecol. 2014;23: 2178–2192.

25. Hewitt G. The genetic legacy of the Quaternary ice ages. Nature. 2000;405: 907–913.

26. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475: 493–496.

27. Hung C-M, Shaner P-JL, Zink RM, Liu W-C, Chu T-C, Huang W-S, et al. Drastic population fluctuations explain the rapid extinction of the passenger pigeon. Proc Natl Acad Sci U S A. 2014;111: 10636–10641.

28. Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, et al. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nat Genet. 2013;45: 67–71.

29. Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. Mol Ecol. 2016;25: 1058–1072.

30. Lloyd GT. Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. Biol J Linn Soc Lond. 2016;118:

131–151.

31. Mild K, Alstrom P. Pipits and Wagtails of Europe, Asia and North America. A&C Black; 2010.

32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10.

33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.

34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.

36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158.

38. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012;3: 35.

39. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15: 356.

40. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers.

Bioinformatics. 2008;24: 1403–1405.

41. Stamatakis A. Using RAxML to Infer Phylogenies. Current Protocols in Bioinformatics. 2015. pp. 6.14.1–6.14.14.

42. Nadachowska-Brzyska K, Li C, Smeds L, Zhang G, Ellegren H. Temporal Dynamics of Avian Populations during Pleistocene Revealed by Whole-Genome Sequences. Curr Biol. 2015;25: 1375–1380.

43. Smeds L, Qvarnström A, Ellegren H. Direct estimate of the rate of germline mutation in a bird. Genome Res. 2016;26: 1211–1218.

44. Ruegg KC, Anderson EC, Paxton KL, Apkenas V, Lao S, Siegel RB, et al. Mapping migration in a songbird using high-resolution genetic markers. Mol Ecol. 2014;23: 5726–5739.

45. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25: 1189–1191.

46. Willing E-M, Dreyer C, van Oosterhout C. Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. PLoS One. 2012;7: e42649.

47. Hartl DL, Clark AG. Principles of Population Genetics. Sinauer Associates Incorporated; 2007.

48. Kozma R, Melsted P, Magnússon KP, Höglund J. Looking into the past - the reaction of three grouse species to climate change over the last million years using whole genome sequences. Mol Ecol. 2016;25: 570–580.

49. Varki A. Comparing the human and chimpanzee genomes: Searching for needles in a haystack. Genome Res. 2005;15: 1746–1758.

50. Clark PU, Mix AC. Ice sheets and sea level of the Last Glacial Maximum. Quat Sci Rev.

2002;21: 1–7.

51. Olsen L, Sveian H, van der Borg K, Bergstram B, Broekmans M. Rapid and rhythmic ice sheet fluctuations in western Scandinavia 15-40 Kya–a review. Polar Res. 2002;21: 235–242.

52. Nosil P, Schluter D. The genes underlying the process of speciation. Trends Ecol Evol. 2011;26: 160–167.

53. Asher G, Schibler U. Crosstalk between components of circadian and metabolic cycles in mammals. Cell Metab. 2011;13: 125–137.

54. Frankham R. Faculty of 1000 evaluation for Contemporary evolution of reproductive isolation and phenotypic divergence in sympatry along a migratory divide [Internet]. F1000 - Post-publication peer review of the biomedical literature. 2010. doi:10.3410/f.1387984.861090

55. Bensch S, Åkesson S, Irwin DE. The use of AFLP to find an informative SNP: genetic differences across a migratory divide in willow warblers. Mol Ecol. 2008;11: 2359–2366.

56. Ruegg KC. The Origin and Maintenance of a Migratory Divide in the Swainson's Thrush (Catharus Ustulatus) and Its Implications for Speciation. 2007.

57. Hobson KA, Kardynal KJ, Van Wilgenburg SL, Albrecht G, Salvadori A, Cadman MD, et al. A Continent-Wide Migratory Divide in North American Breeding Barn Swallows (*Hirundo rustica*). PLoS One. 2015;10: e0129340.

58. Olivares C, Solano F. New insights into the active site structure and catalytic mechanism of tyrosinase and its related proteins. Pigment Cell Melanoma Res. 2009;22: 750–760.

59. Huang Y-H, Lee T-H, Chan K-J, Hsu F-L, Wu Y-C, Lee M-H. Anemonin is a natural bioactive compound that can regulate tyrosinase-related proteins and mRNA in human melanocytes. J Dermatol Sci. 2008;49: 115–123.

60. Xu Y, Zhang X-H, Pang Y-Z. Association of Tyrosinase (TYR) and Tyrosinase-related Protein 1 (TYRP1) with Melanic Plumage Color in Korean Quails (Coturnix coturnix). Asian-australas J Anim Sci. 2013;26: 1518–1522.

61. Manceau M, Domingues VS, Mallarino R, Hoekstra HE. The developmental role of agouti in color pattern evolution. Science. 2011;331: 1062–1065.

62. Poelstra JW, Vijay N, Hoeppner MP, Wolf JBW. Transcriptomics of colour patterning and coloration shifts in crows. Mol Ecol. 2015;24: 4617–4628.

63. Price TD. Phenotypic plasticity, sexual selection and the evolution of colour patterns. J Exp Biol. 2006;209: 2368–2376.

64. Campagna L, Repenning M, Silveira LF, Fontana CS, Tubaro PL, Lovette IJ. Repeated divergent selection on pigmentation genes in a rapid finch radiation. Science Advances. 2017;3: e1602404.

65. Kirkpatrick M, Hall DW. Sexual selection and sex linkage. Evolution. 2004;58: 683–691.

66. Vicoso B, Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. Nat Rev Genet. 2006;7: 645–653.

67. Wright AE, Harrison PW, Zimmer F, Montgomery SH, Pointer MA, Mank JE. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. Mol Ecol. 2015;24: 1218–1235.

68. Wagner CE, McCune AR, Lovette IJ. Recent speciation between sympatric Tanganyikan cichlid colour morphs. Mol Ecol. 2012;21: 3283–3292.

69. Toews DPL, Taylor SA, Vallender R, Brelsford A, Butcher BG, Messer PW, et al. Plumage genes and little else distinguish the genomes of hybridizing warblers. Curr Biol. 2016;26: 2313–2318.

Table 1. Sequencing and alignment summary statistics.

| Species | ID | Total reads (P1+P2) | % reads aligned to reference | Mean (sd) depth Q>20 | % chimeras | % breadth cov. (DP>0) | % breath cov. (DP>5) | % hetero-zygosity |
|---|---|---|---|---|---|---|---|---|
| *M. a. alba* | alb045 | 94915300 | 51.7 | 9.9 (30.1) | 2.7 | 55 | 50.7 | 0.47 |
| *M. a. alba** | alb031 | 56234580 | 52.0 | 5.7 (19.4) | 1.5 | 54.1 | 35.5 | 0.45 |
| *M. a. baicalensis* | alb074 | 94163802 | 51.1 | 9.6 (32.3) | 2.7 | 55 | 50.4 | 0.47 |
| *M. a. baicalensis* | alb030 | 105148514 | 51.9 | 10.7 (33.8) | 2.8 | 55.1 | 51.5 | 0.49 |
| *M. a. lugens* | alb041 | 117417412 | 51.1 | 11.6 (37.4) | 2.8 | 55.2 | 52.3 | 0.46 |
| *M. a. lugens** | alb057 | 183433508 | 52.1 | 46.7 (70.4) | 1.8 | 55.6 | 54.9 | 0.41 |
| *M. a. ocularis* | alb010 | 96304060 | 51.7 | 9.8 (32.3) | 2.7 | 55 | 50.7 | 0.47 |
| *M. a. ocularis* | alb039 | 87166124 | 51.5 | 8.9 (28.6) | 2.6 | 54.9 | 49.4 | 0.47 |
| *M. a. personata* | alb027 | 106647088 | 51.6 | 11.2 (27.3) | 2.1 | 55.1 | 51.8 | 0.47 |
| *M. a. personata** | alb008 | 52519228 | 51.8 | 5.3 (23.2) | 1.9 | 53.9 | 32.1 | 0.45 |
| *M. f. feldegg* | fl046 | 103658294 | 51.1 | 10.3 (22.9) | 2.2 | 55.2 | 51.5 | 0.95 |
| *M. f. feldegg** | fl004 | 52689618 | 51.7 | 5.2 (26.4) | 2.3 | 54.2 | 31.5 | 0.91 |
| *M. f. flava* | fl076 | 109692042 | 51.0 | 10.7 (39.4) | 3.0 | 55.2 | 51.8 | 0.97 |
| *M. f. flava / beema* | fl021 | 109960090 | 51.6 | 10.7 (43.8) | 3.2 | 55.2 | 51.7 | 0.93 |
| *M. f. flava** | fl068 | 105420824 | 50.1 | 9.7 (21.5) | 1.7 | 55.2 | 49.9 | 0.97 |
| *M. f. lutea* | fl060 | 34497972 | 51.5 | 3.6 (18.1) | 2.8 | 51.4 | 18.7 | 0.84 |
| *M. f. taivana* | fl035 | 127173220 | 51.0 | 12.1 (44.1) | 3.1 | 55.4 | 52.9 | 0.86 |
| *M. f. taivana** | fl074 | 35374233 | 51.6 | 7.2 (27.6) | 2.0 | 50 | 40 | 0.81 |
| *M. f. tschutschensis* | fl010 | 113519472 | 50.6 | 10.9 (34.6) | 2.5 | 55.3 | 52.1 | 0.9 |
| *M. f. tschutschensis* | fl029 | 137937380 | 51.3 | 13.0 (47.3) | 3.1 | 55.4 | 53.2 | 0.91 |

* denotes individuals sequenced at Brigham Young University, all others sequenced at UC Berkeley.

Table 2. Genes of interest involved in plumage pigment (left) and circadian rhythm (right) pathways.

| Plumage | | Migration | |
|---|---|---|---|
| **Gene** | **Description** | **Gene** | **Description** |
| ANKRD27 | Uncharacterized protein | AANAT | aralkylamine N-acetyltransferase |
| ANXA6 | Annexin A6 | ADCYAP1 | adenylate cyclase activating polypeptide 1 |
| AP1G1 | Uncharacterized protein | ARNTL | aryl hydrocarbon receptor nuclear translocator like |
| AP1M1 | Uncharacterized protein | ARNTL2 | aryl hydrocarbon receptor nuclear translocator like 2 |
| ATP6V0A1 | V-type proton ATPase 116 kDa subunit a isoform 1 | CIPC | CLOCK interacting pacemaker |
| BCO1 | Beta,beta-carotene 15,15'-dioxygenase | CLOCK | clock circadian regulator |
| BLOC1S6 | Biogenesis of lysosome-related organelles complex 1 subunit 6 | HSPA5 | heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa) |
| BSG | Basigin | HSPA8 | heat shock 70kDa protein 8 |
| CITED2 | Cited2/melanocyte specific gene-related gene 1 MRG1 | NEK2 | NIMA-related kinase 2 |
| DDT | D-dopachrome decarboxylase | NFIL3 | nuclear factor, interleukin 3 regulated |
| EDNRB | Uncharacterized protein | NPAS2 | neuronal PAS domain protein 2 |
| GPR143 | Uncharacterized protein | PER2 | period circadian clock 2 |
| HPS1 | Uncharacterized protein | PER3 | period circadian clock 3 |
| HPS4 | Uncharacterized protein | SLC1A3 | solute carrier family 1 (glial high affinity glutamate transporter), member 3 |
| HPS6 | Uncharacterized protein | SLC2A1 | solute carrier family 2 (facilitated glucose transporter), member 1 |
| HSP90AA1 | Heat shock protein HSP 90-alpha | SLC2A10 | solute carrier family 2 (facilitated glucose transporter), member 10 |
| HSP90AB1 | Heat shock cognate protein HSP 90-beta | SLC2A12 | solute carrier family 2 member 12 |
| HTR7 | 5-hydroxytryptamine receptor 7 transcript variant 1 | | |
| ITGB1 | Integrin beta-1 | | |
| MC1R | Melanocyte-stimulating hormone receptor | | |
| MC4R | Uncharacterized protein | | |
| METTL4 | Uncharacterized protein | | |
| MITF | Uncharacterized protein | | |
| PPIB | Peptidyl-prolyl cis-trans isomerase B | | |
| RAB17 | Uncharacterized protein | | |
| RAB38 | Uncharacterized protein | | |
| SLC24A5 | Uncharacterized protein | | |
| TFRC | Transferrin receptor protein 1 | | |
| TMEM33 | Uncharacterized protein | | |
| TYR | Tyrosinase | | |
| WNT5A | Protein Wnt | | |

Table 3. Pairwise weighted Weir & Cockerham's Fst.

| | ***alba*** | ***baicalensis*** | ***lugens*** | ***ocularis*** |
|---|---|---|---|---|
| ***baicalensis*** | 0.1842 | | | |
| ***lugens*** | 0.2298 | 0.1975 | | |
| ***ocularis*** | 0.2034 | 0.1741 | 0.1843 | |
| ***personata*** | 0.1963 | 0.1789 | 0.2223 | 0.1963 |
| | ***feldegg*** | ***flava*** | ***lutea*** | ***taivana*** |
| ***flava*** | 0.1399 | | | |
| ***lutea*** | 0.3224 | 0.2398 | | |
| ***taivana*** | 0.3443 | 0.2984 | 0.4549 | |
| ***tschutschensis*** | 0.3176 | 0.2746 | 0.4189 | 0.2344 |

Table 4. Percentage of fixed SNPs, full dataset (Q > 20, depth > 5x, 0% missing across species).

| | ***alba*** | ***baicalensis*** | ***lugens*** | ***ocularis*** |
|---|---|---|---|---|
| ***baicalensis*** | 0.009% | | | |
| ***lugens*** | 0.019% | 0.015% | | |
| ***ocularis*** | 0.011% | 0.008% | 0.011% | |
| ***personata*** | 0.01% | 0.001% | 0.019% | 0.011% |
| | ***feldegg*** | ***flava*** | ***lutea*** | ***taivana*** |
| ***flava*** | 0.012% | | | |
| ***lutea*** | 0.064% | 0.063% | | |
| ***taivana*** | 0.142% | 0.142% | 0.215% | |
| ***tschutschensis*** | 0.135% | 0.135% | 0.197% | 0.034% |

Figure 1. PCA of discrete plumage characteristics. Top: *M. alba*, Bottom: *M. flava*. Subspecies not included in the study are in black.

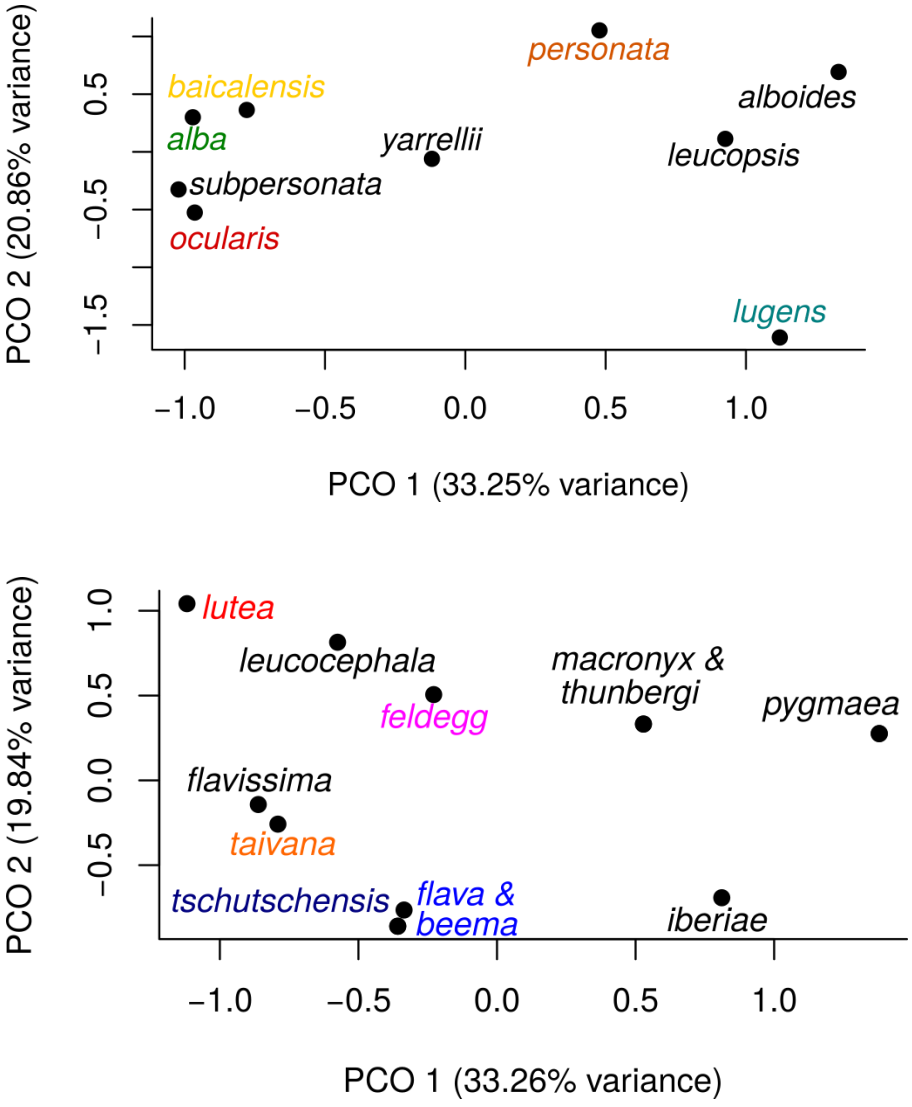


Figure 1. Sampling localities and PCA using ~37K SNPs, 0% missing data.

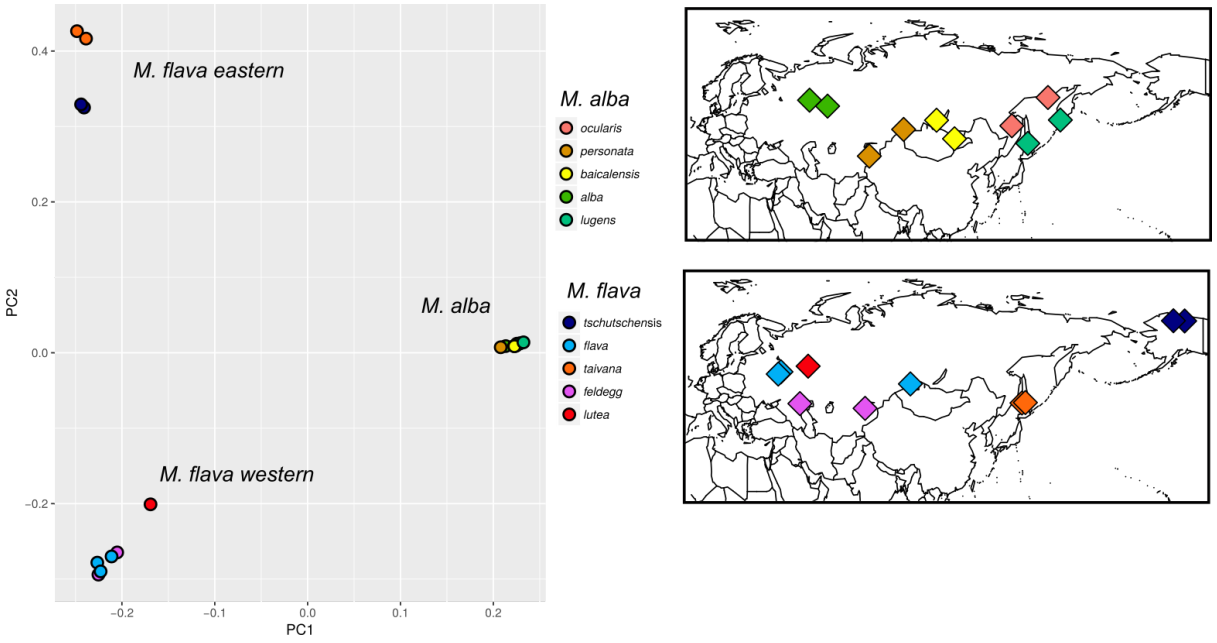


Figure 3. RAxML tree using ~10K concatenated SNPs. Bootstrap supported indicated at nodes, circles represent bootstrap > 95%. Tips in black are *M. flava* subspecies, whereas tips in blue are *M. alba* subspecies.

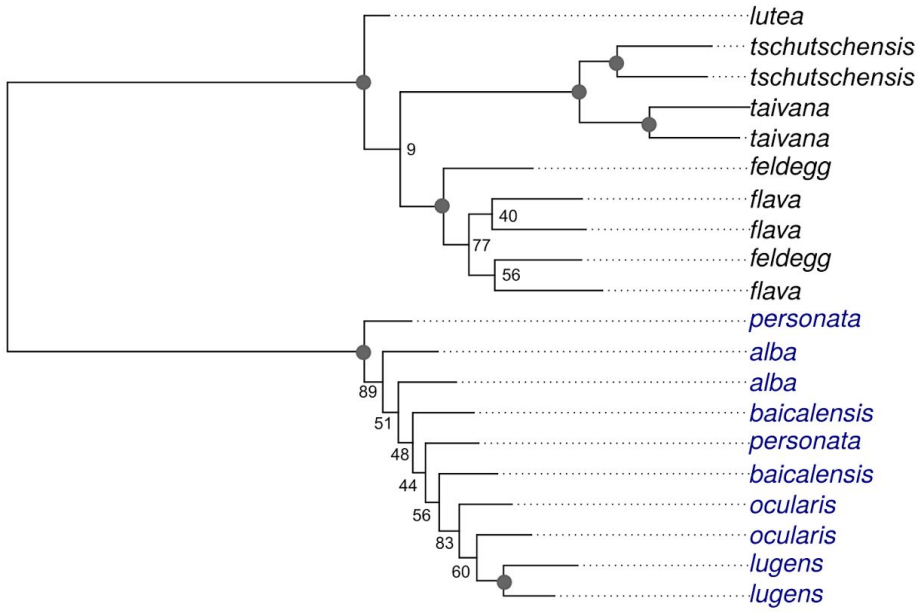


Figure 4. Demographic history of *M. flava* subspecies using generation time (g) = 1, and mutation rate (u) = 2.3 x 10^-9. Dark lines show results of all data analysis, thin lines show results of bootstrap replicates. According to Chapter 1, *tschutschensis* and *taivana* make up the eastern population, while *beema, flava,* and *feldegg* make up the western population.

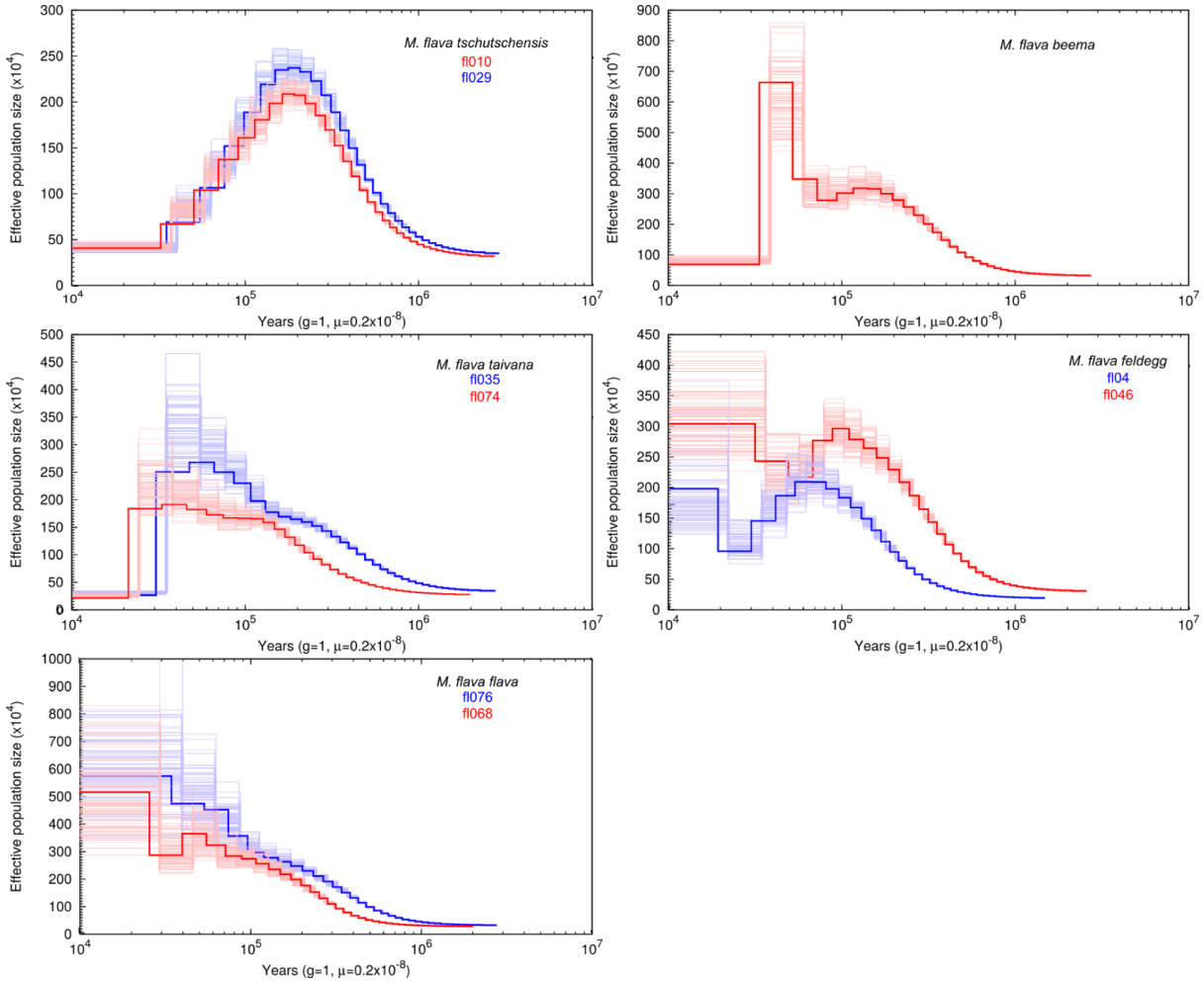


Figure 5. Demographic history of *M. alba* subspecies using generation time (g) = 1, and mutation rate (u) = 2.3 x 10^-9. Dark lines show results of all data analysis, thin lines show results of bootstrap replicates. According to the results of Chapter 1, the top row (*ocularis* and *lugens*) below to the eastern population, whereas the bottom two rows (*baicalensis*, *personata*, and *alba*) belong to the western population.

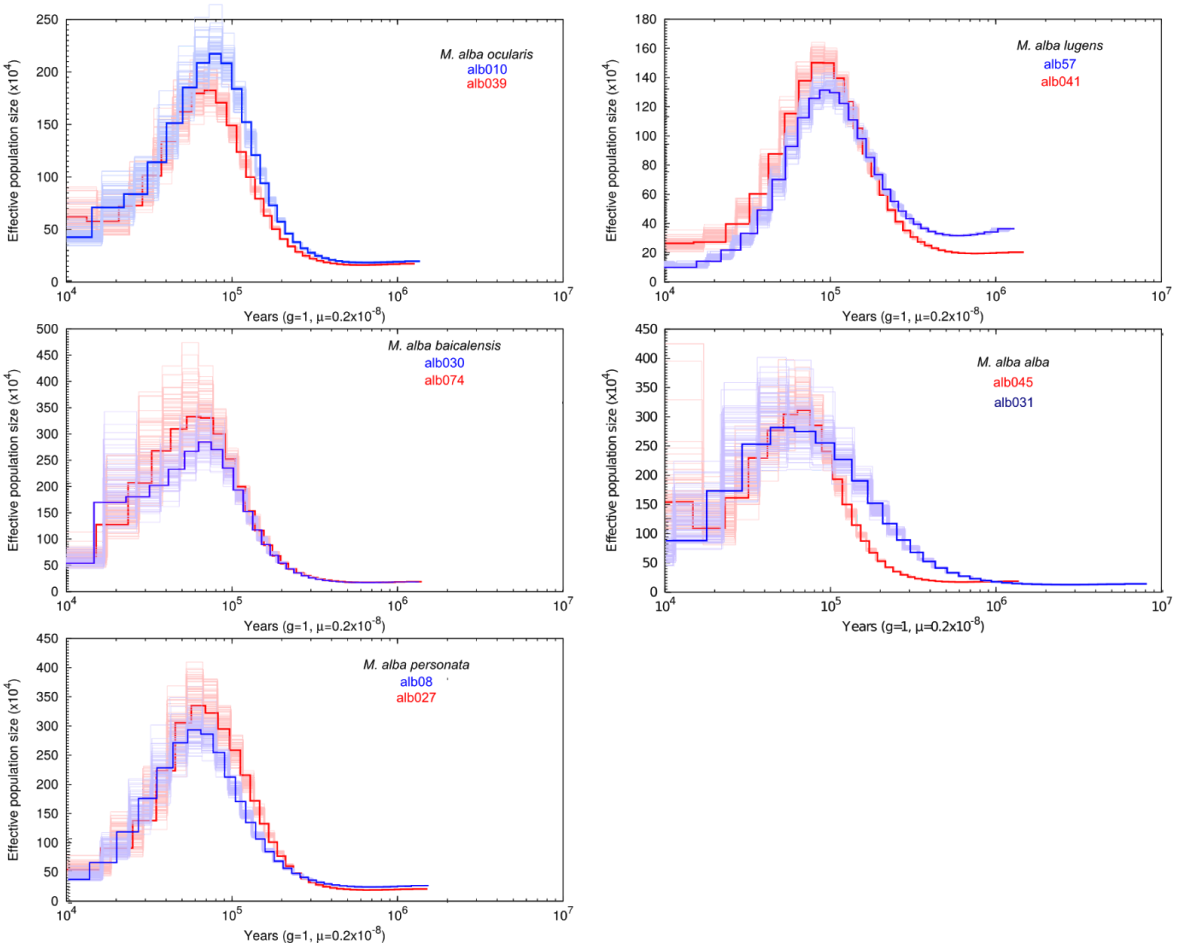


Figure 6. Demographic history of *M. alba* and *M. flava.* One representative of each subspecies is shown.

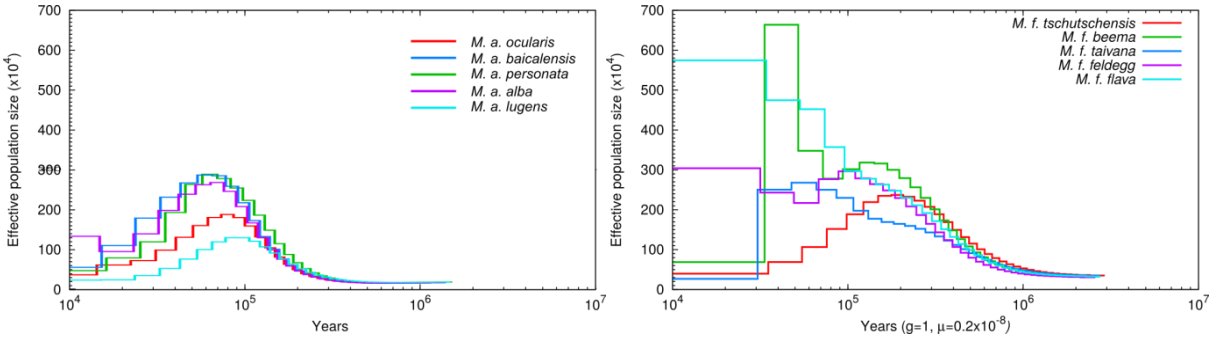


Figure 7. Sliding window (10 kbp) analysis of fixed SNPs in the *M. alba* species complex: top) eastern (*ocularis, lugens*) vs. western (*alba, baicalensis, personata*), middle) intra-eastern, bottom) intra-western. Chromosomal position reflects mapping of scaffolds to zebra finch genome.

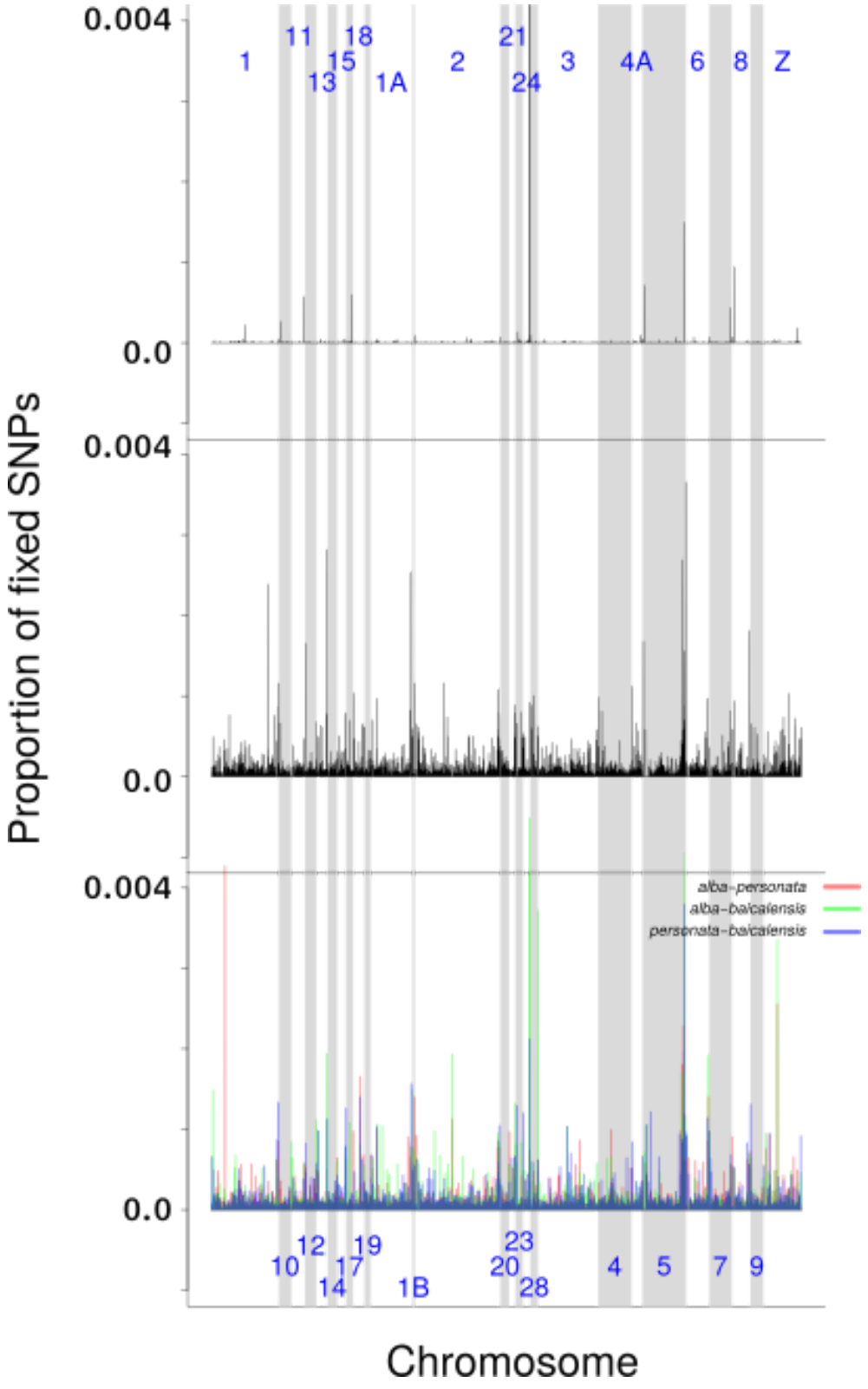



Figure 8. Sliding window (10 kbp) analysis of fixed SNPs in the *M. flava* species complex: top)

eastern (*beema, lutea, flava, feldegg*) vs. western (*tschutschensis, taivana*), middle) intra-eastern (*tschutschensis* (NE) vs. *taivana* (SE)), bottom) intra-western (*flava, feldegg*). Chromosomal position reflects mapping of scaffolds to zebra finch genome.

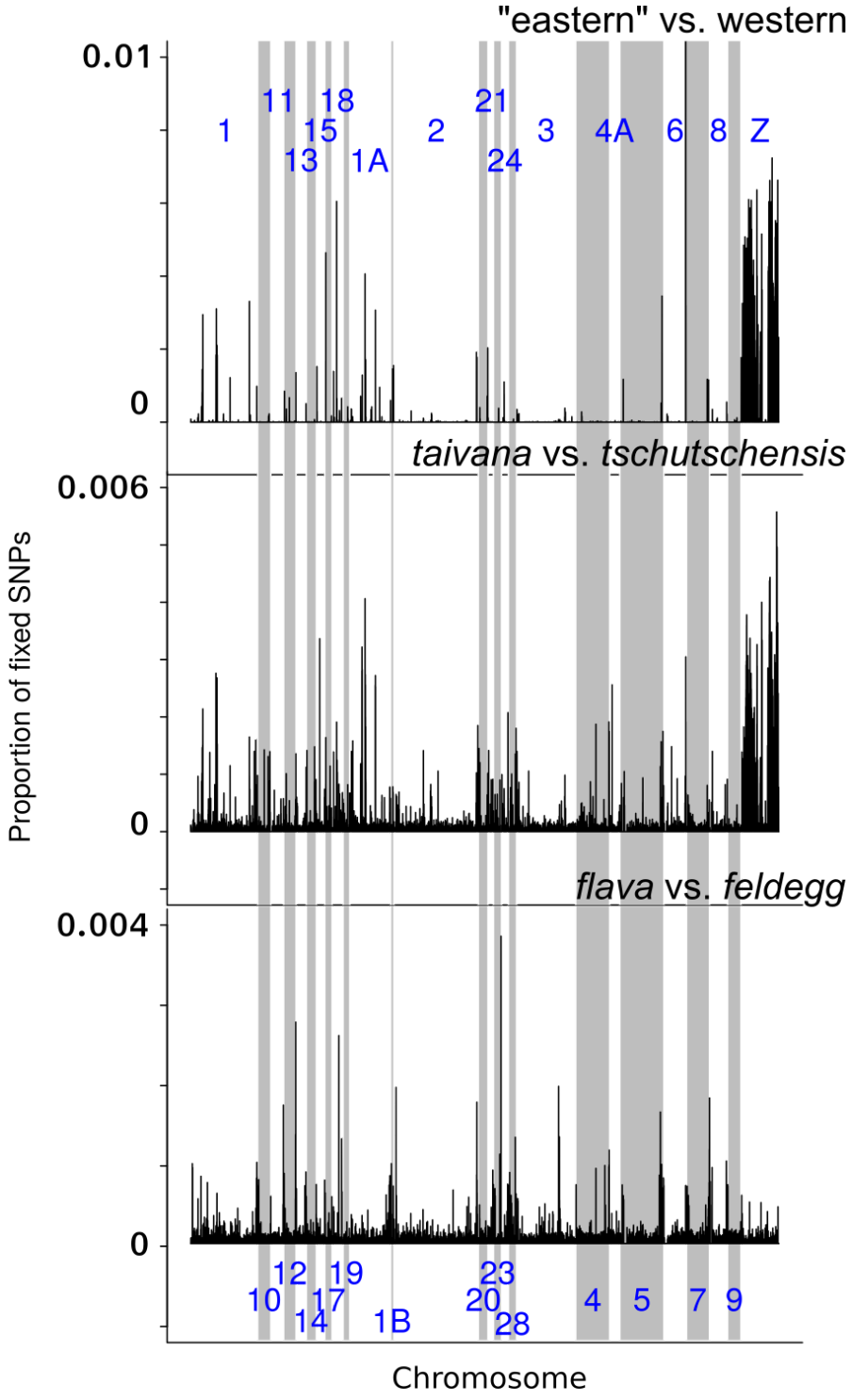


Figure 9. Fst sliding window (300 kbp window, 30 kbp step) analysis of *M. flava* "eastern" and western populations. Horizontal red line represents the genome-wide Fst value (0.165).