Evolution, 2025, **79**(**12**), 2612–2628
https://doi.org/10.1093/evolut/qpaf202
Advance access publication 9 October 2025
Original Article

Editor's choice

How accurate is genomic prediction across wild populations?

Kenneth Aase[1,2], **Hamish A. Burnett**[2,3], **Henrik Jensen**[2,3], **Stefanie Muff**[1,2]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[2]The Gjærevoll Centre, Norwegian University of Science and Technology, Trondheim, Norway
[3]Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

Corresponding author: Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz' vei 1, 7034 Trondheim, Trøndelag, Norway. Email: kenneth.aase@ntnu.no

Abstract

Evolutionary ecology seeks to understand causes and consequences of evolutionary changes across time and space, and genomic data present novel opportunities to investigate these processes. Genomic prediction—predicting individual genetic values from high-density marker data—has revolutionized breeding programs and medical genetics. In wild populations, however, genomic prediction has been used in comparatively few studies, and largely *within* populations. Applications that instead operate *across* populations could answer questions related to spatially varying evolutionary processes, such as local adaptation. A severe challenge for across-population genomic prediction, however, is the decrease in accuracy when training models on data from one population and predicting genetic values in another. Here, we applied genomic prediction across wild house sparrow populations and compared the accuracy to within-population models. We also highlighted limitations of the current theory for genomic prediction accuracy, and sought to provide a mechanistic understanding of the across-population accuracy by relating it to several population-differentiation measures. Predictions across populations were generally less accurate and more variable than within populations, and across-population accuracy covaried with some population-differentiation metrics. Our results underline the necessity of understanding the mechanisms governing genomic prediction accuracy, and of developing methods that exploit genomic data in novel ways.

Keywords: accuracy, across-population, genomic prediction, house sparrow, population-differentiation, wild populations

Introduction

Quantifying the genetic basis of a given phenotypic trait is useful for understanding how evolutionary processes act in nature (Charmantier et al., 2014; Kruuk et al., 2008; Lynch & Walsh, 1998; Walsh & Lynch, 2018). In wild populations, the prediction of an individual's additive genetic value (also known as breeding value), which is the heritable contribution of the individual's genome to its phenotype (Hill, 2014), can help us track adaptive and non-adaptive evolutionary changes across time, detect local adaptation to spatially varying environmental conditions, investigate selection processes, make inferences about unmeasured or unexpressed phenotypes in genotyped individuals, and compare how much individuals' genotypes contribute to trait variation relative to environmental conditions (Hunter et al., 2022; Jensen et al., 2014). Traditionally, genetic values for complex traits have been predicted using relatedness information from multi-generational pedigrees (Hadfield et al., 2010; Henderson, 1984; Kruuk, 2004), but thanks to advances in genotyping technologies, we are now less reliant on the labor- and time-consuming task of constructing pedigrees from long-term study systems. Methods that instead rely on genomic data, usually individual genotype information on high-density genome-wide sets of single-nucleotide polymorphism (SNP) markers, to predict genetic values are known as *genomic prediction* (GP).

The idea of GP originated in animal breeding (Meuwissen et al., 2001), where the goal is artificial selection on profitable traits. By exploiting genomic data from the animals, the genetic values can be predicted already in early life stages, speeding up the selection process by eliminating the need to wait for a trait to be expressed in the individual itself or its descendants. Additionally, GP is expected to more accurately predict genetic values compared to pedigree-based methods (Gienapp et al., 2017). Due to these factors and the decreasing cost of SNP genotyping, GP is now widely used in animal and plant breeding (Crossa et al., 2017; Hickey et al., 2017; Meuwissen et al., 2016). Furthermore, medical geneticists are developing *polygenic risk scoring*, a variation of GP that shows promise for use in personalized medicine to detect an individual's genetic risk for complex diseases (Choi et al., 2020; Khera et al., 2018; Wray et al., 2019).

In wild populations, application of GP is still in its infancy (McGaugh et al., 2021), possibly due to additional challenges present in wild study systems compared to breeding or medical genetics. Firstly, high-density genomic data sets are less common, and sample sizes are generally smaller

Received June 17, 2025; revisions received September 22, 2025; accepted September 29, 2025
Associate Editor: Henrique Teotonio; Handling Editor: Jason Wolf

than in the aforementioned fields because of inherent difficulties and higher costs of sampling wild animals. Second, the effective population sizes N_e are often larger in wild systems than in agriculture, making it more challenging to obtain accurate and precise estimates of additive genetic effects (Gienapp et al., 2019). A third important challenge is the need to account for environmental heterogeneity and population structure (Aase et al., 2022; Kruuk & Hadfield, 2007), both of which are common in wild populations, but outside the researcher's control. Given these challenges, it is not surprising that only a few GP studies have been performed in wild animal populations, for example, in great tits (*Parus major*, see Gienapp et al., 2019; Lindner et al., 2023, 2024; Verhagen et al., 2019), Soay sheep (*Ovis aries*, see Ashraf et al., 2022; Hunter et al., 2022; Vahedi et al., 2023), red deer (*Cervus elaphus*, see Gauzere et al., 2023), three-spine stickleback (*Gasterosteus aculeatus*, see Strickland et al., 2024), and house sparrows (*Passer domesticus*, see Aspheim et al., 2024). A shared methodological characteristic among these studies was that they all used GP to predict the genetic values of individuals sampled from the same (meta-)population as the individuals used to train (i.e., fit) the model. However, many potential applications of GP are reliant on prediction from one population into *another* population, as was done by Bosse et al. (2017).

In light of the high cost and aforementioned difficulty of sampling wild populations, borrowing information from phenotyped populations to make inferences about non-phenotyped individuals in other populations would be a very useful application of GP. *Across-population* GP here refers to models and applications in which the phenotyped and genotyped individuals used to fit the model and the unmeasured, but genotyped, individuals for which we predict genetic values are sampled from two different populations. With reliable across-population GP methods, one would be able to train a GP model for hard-to-measure phenotypes in one population, and only need to genotype another population to make inferences about this population's genetic values, greatly expanding the number of populations where we are able to answer the aforementioned evolutionary and quantitative genetic questions. Across-population GP could also be especially relevant in conservation programs that aim to use captive breeding and translocation to strengthen existing natural populations (e.g., Sauve et al., 2022), where the predicted fitness of an individual in another population is of interest. Similarly, GP may also be useful across temporal, as opposed to spatial, distances between the training and test samples (Habier et al., 2007). Contemporary genotype and phenotype data coupled with ancient genetic material could allow for inferences about ancestors that are temporally distant and provide information on the rate and direction of any evolutionary change that has occurred within the population over time (e.g., Berens et al., 2017; Cox et al., 2019; Marciniak et al., 2022).

A major problem, known from both agricultural and medical applications, is that GP across populations (i.e., breeds or ethnicities) suffers from seriously degraded prediction accuracy compared to within-population GP (Habier et al., 2007; Hayes et al., 2009a; Lupi et al., 2024; Martin et al., 2017, 2019; Rio et al., 2021). Similar issues have also been discussed in the context of using polygenic scores in paleogenetics, where the target population is ancient, and thus very temporally distant from contemporary training data (Carlson et al., 2022; Irving-Pease et al., 2021). The problem of reduced prediction accuracy has particularly been recognized in medical genetics, where early attempts to apply polygenic scores derived from one population to others sparked significant debate about both methodological validity and potential misuse (Berg & Coop, 2014; Novembre & Barton, 2018). This ongoing challenge in several fields highlights the critical importance of understanding the fundamental mechanisms underpinning across-population GP.

In general, a theoretical understanding of expected GP accuracy in a broad range of scenarios—including both within- and across-population predictions—would be essential. Some previous work has developed theoretical frameworks and mathematical formulas for forecasting GP accuracy in given within-population scenarios (Daetwyler et al., 2008; Dekkers et al., 2021; Goddard et al., 2011; Lee et al., 2017), and how much the accuracy is expected to degrade when predicting across populations (Ding et al., 2023; Wang et al., 2020; Wientjes et al., 2015). When planning studies, for example, such deterministic formulas can help decide on the required amount of field sampling and genotyping, and can allow one to assess a priori whether a planned experiment or study has the power to deliver the desired insight. Additionally, various explanations for the mechanisms behind the lowered across-population GP accuracy have been proposed, including population differences in allele frequencies or linkage disequilibrium (LD) patterns, lack of family relationships, or differences in allele effects caused by non-additive genetic effects and/or genotype-by-environment interactions (Hou et al., 2023; Wang et al., 2020; Zhong et al., 2009). However, as we will discuss, the formulas for expected accuracy suffer from various deficiencies, especially when applied across populations, and the problem of understanding across-population GP accuracy still needs further investigation.

Wild systems differ from the populations used in animal breeding and human genetics in many ways. Thus, results from those fields are not directly transferable to evolutionary ecology, so how well across-population GP works in wild populations has not previously been examined. Furthermore, the challenges with across-population GP are not well known within evolutionary ecology. Bosse et al. (2017), the only study to our knowledge that performed across-population GP in the wild, did not mention the problems with reduced accuracy. In order to improve our understanding of how well standard methods for GP work when used across populations, we here compared the accuracy of within- and across-population GP for morphological traits of wild house sparrow populations. The island structure of the study populations, along with the relatively large sample sizes from this long-term study system, makes this data set uniquely suited to explore any challenges of across-population GP. We compared the performance of GP within and across populations, and explored whether various measures of genetic differentiation between populations affected the across-population accuracy. Additionally, we discussed various aspects of GP accuracy, including proper scaling methods for accuracy and how to deal with repeated measurements, and we demonstrated that existing formulas for expected accuracy are inappropriate for across-population GP.

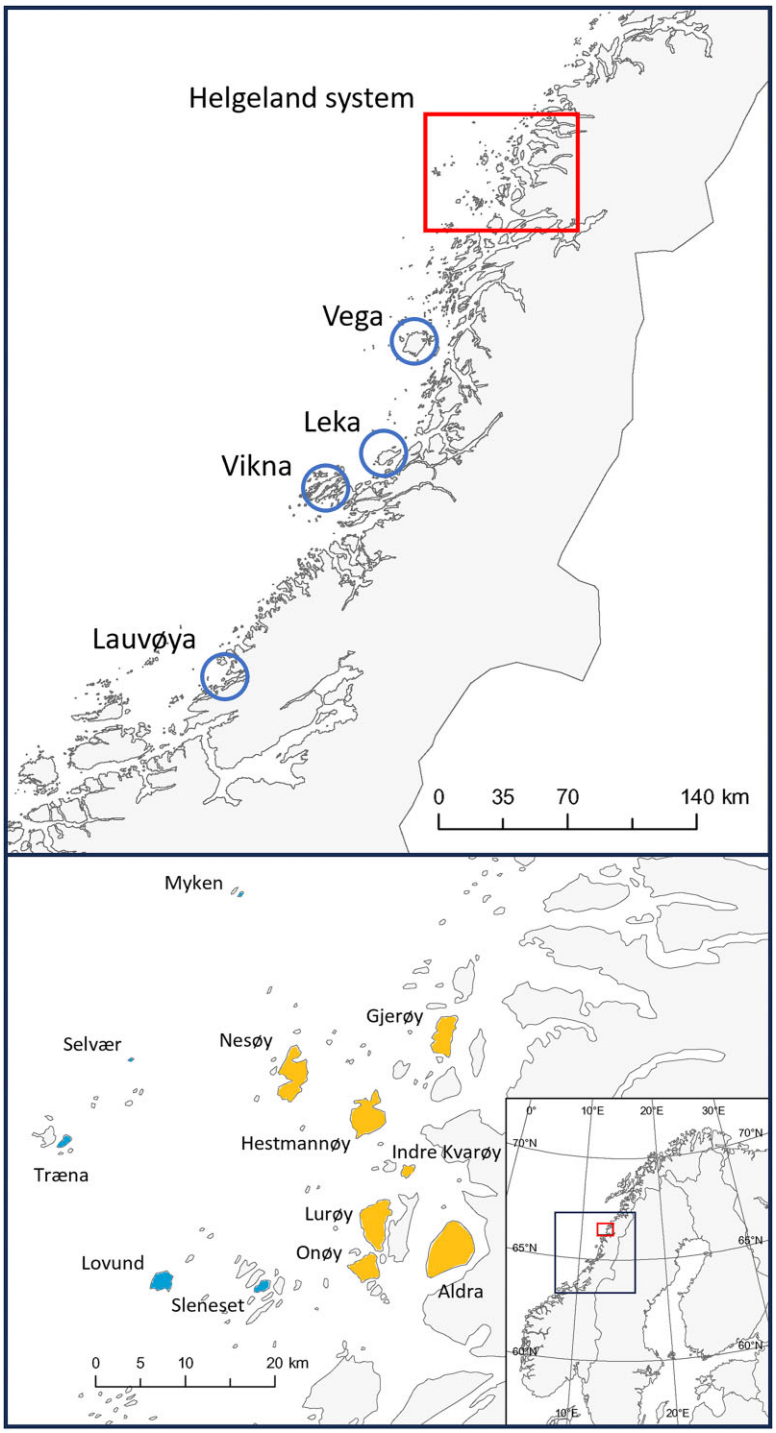

Figure 1. Maps showing the locations of the various island populations in the study. In the top panel, the Helgeland system is indicated with a red square, while the blue circles indicate the southern islands of Vega, Leka, Vikna, and Lauvøya. The bottom panel shows the Helgeland system, with farm islands colored yellow and non-farm islands colored blue. The map in the bottom right shows the location of the other two maps, zoomed out to the scale of Norway.

Methods and materials

House sparrow data

The data underlying all GP analyses presented here stem from long-term individual-based house sparrow study populations located on several islands along the coast of Norway (see Figure 1 and Jensen et al., 2013), including 12 islands in an insular meta-population on the coast of Helgeland (Ranke et al., 2021; Sæther et al., 1999) and 4 islands 75–295 km south of this meta-population (Kvalnes et al., 2017; Nafstad et al., 2023; Ranke et al., 2020). The long-term nature of the study (most populations have been studied since 1993), along with high capture rates, relatively low levels of natal dispersal, and virtually no breeding dispersal (Ranke et al., 2021), makes it attractive for investigating evolutionary ecological hypotheses. When the spar-

rows are first captured, they are marked with a numbered metal ring and three colored plastic rings, which uniquely identify them in future recaptures and re-sightings. Birds first ringed as nestlings or fledged juveniles in the summer were known to have hatched that year. Because of the high ringing and recapture rate in our study populations, birds first captured in the autumn were assumed to have hatched the same year as they were first captured, while birds first captured as adults from January to August were assumed to have hatched the previous year (Araya-Ajoy et al., 2021; Niskanen et al., 2020). The study populations have been subject to an extensive pheno- and genotyping effort, with a total of over 12,000 house sparrows genotyped on either a custom Axiom 200K SNP array (Lundregan et al., 2018; Niskanen et al., 2020) or a cross-compatible Axiom 70K SNP array containing a subset of the SNPs on the 200K array with identical probe design (Burnett et al., 2025). Within chromosomes, SNPs included on the arrays were spaced at an average distance of 8–9 kb. A subset of these genotyped individuals was recorded as adults and measured for various morphological traits, including body mass (to the nearest 0.1 g), tarsus length (to the nearest 0.01 mm), and wing length (to the nearest millimeter), with repeated measurements over several years available for many of the individuals and measurements adjusted for any fieldworker differences (see, e.g., Niskanen et al., 2020).

The hierarchical geographic island structure in the study systems makes the house sparrow data ideal for investigating the accuracy of across-population GP. We leverage this natural structure to investigate a range of situations relevant for wild populations in general. As explained below, we organize the data into three scenarios with increasingly stronger genetic population structure (Jensen et al., 2013; Nafstad et al., 2023; Ranke et al., 2024): the Helgeland meta-population (*Scenario 1*, islands connected through dispersal), the southern islands (*Scenario 2*, islands connected through translocations), and a merged scenario (*Scenario 3*) where we use all the available data from Scenarios 1 and 2. Overall sample sizes for each trait in each scenario are shown in Table 1, though each GP model was trained on a subset of these samples, as explained below.

Scenario 1: Helgeland

The first scenario is based on the data from the Helgeland system, a meta-population located in an archipelago off the coast of Northern Norway. We used data collected on the islands in the archipelago, which are named in the map in the lower panel in Figure 1. Although the house sparrow is generally a sedentary bird species, the island populations in the Helgeland archipelago are connected by the sparrows that disperse between the islands during the autumn in the year they are born (Ranke et al., 2021; Saatoglu et al., 2021). The islands are thus genetically connected by spatially varying degrees of dispersal, resulting in varying levels of genetic differentiation between them. Notably, the islands can broadly be categorized into "farm" islands, where the sparrows usually nest in the barns of dairy farms, or "non-farm" islands, where the sparrows live in local people's gardens where their environment is generally harsher outside the breeding season. Previous studies have found that individuals from these two island types are phenotypically and genetically differentiated (Holand et al., 2011; Jensen et al., 2013; Saatoglu et al., 2024), and in particular that they differ in both the means and variances of the genetic values for various morphological traits (Aase et al., 2022; Muff et al., 2019).

Scenario 2: Southern

The second scenario corresponds to four "farm" islands located south of the Helgeland meta-population (Vega, Leka, Vikna, and Lauvøya; see Figure 1, upper panel). Due to the distances between these four islands, there is virtually no natural dispersal between them (Ranke et al., 2024). However, some of the southern island populations have been connected by past translocation experiments, in which individuals were moved between the islands (for details, see Kvalnes et al., 2017; Nafstad et al., 2023; Ranke et al., 2020). In other words, Scenario 2 consists of separate, but partially connected island populations, and thus has stronger genetic population structure than the Helgeland scenario (Jensen et al., 2013).

Scenario 3: Merged

In the third scenario, denoted *merged*, we used all the data from the first two scenarios (Helgeland system and southern islands) simultaneously. Estimates of genetic differentiation suggest low dispersal rates between the Helgeland system and the southern islands (Jensen et al., 2013), which is supported by only a handful of dispersal events recorded across more than 20 years by recapture and re-sighting of thousands of ringed house sparrows (Ranke et al., 2024). Thus, in practice, the southern islands are wholly separate populations from the Helgeland meta-population, so the data in Scenario 3 have a higher hierarchical level of structure than Scenarios 1 and 2.

Statistical methods

Models for genomic prediction

All GP models were formulated as mixed models with the aforementioned morphological traits (body mass, tarsus length, or wing length) as continuous responses, given as

$$y_{ijk} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + g_i + z_{\text{hatch-year},i} + z_{\text{island},ij} + z_{\text{id},i} + z_{\text{session},ij} + z_{\varepsilon,ijk}\,, \quad (1)$$

where y_{ijk} is the kth phenotypic measurement of individual i during *measurement session* j (see below), $\boldsymbol{x}_{ij}$ is a vector of fixed covariates with corresponding fixed effects vector $\boldsymbol{\beta}$, encoding for an intercept term, a binary variable for phenotypic sex (0 = female, 1 = male), a categorical variable for the month of measurement (from January = 1 through November = 11), and a quantitative variable for age in years (i.e., the difference between measurement year and hatch year; range 1–11). The other terms in equation 1 are random effects, with g_i being a structured Gaussian random effect representing the genetic value of individual i, and the remaining terms being independent unstructured Gaussian random effects accounting for various confounding effects. The random effect $z_{\text{hatch-year},i}$ accounted for the effect of individual i's hatch year (range 1992–2020), $z_{\text{island},ij}$ accounted for the effect of island of measurement (i.e., one of the islands named on Figure 1), and $z_{\text{id},i}$ accounted for remaining permanent environmental effects specific to individual i. Following Ponzi et al. (2018), we defined a measurement session to be the set of measurements of a given individual made on the same day, and included the measurement session-specific random effect $z_{\text{session},ij}$. Since the birds experience

Table 1. The number of unique individuals that were genotyped and measured for a given phenotype.

Trait	Scenario		
	1: Helgeland	2: Southern	3: Merged
Body mass	3,456 (7,977)	2,254 (4,124)	5,710 (12,101)
Tarsus length	3,445 (8,067)	2,255 (4,171)	5,700 (12,238)
Wing length	3,424 (7,997)	2,254 (4,151)	5,678 (12,148)

Note. The numbers in the parentheses are the total number of measurements, which is higher than the number of unique individuals due to repeated measurements.

essentially the same environment during a measurement session, $z_{\text{session},ij}$ thus accounted for all residual variation except measurement error, which is captured by $z_{\varepsilon,ijk}$. We fitted the mixed models in a Bayesian framework using `R-INLA` (Rue et al., 2009), and assigned independent $\mathsf{N}\left(0, 10^3\right)$ priors to all entries in $\boldsymbol{\beta}$. For all random effect variances, we assigned independent penalized complexity priors $\text{PC}\left(\hat{\sigma}/\sqrt{2}, 0.05\right)$, where $\hat{\sigma}$ denotes the sample standard deviation of the respective model's response in the training set (Simpson et al., 2017). In other words, we ascribed a prior probability of 5% that a given random effect would explain more than half the sample variance.

We modeled the genetic value g_i using a genomic animal model formulation, that is, we assumed that the vector $\boldsymbol{g} = \left(g_1, \ldots, g_{N_{\text{ind}}}\right)^\top$ follows a normal distribution $\mathsf{N}\left(\mathbf{0}, \sigma^2_{\text{G}}\mathbf{G}\right)$, where N_{ind} is the number of individuals, σ^2_{G} is the additive genetic variance (often denoted V_A), and $\mathbf{G}$ is an $N_{\text{ind}} \times N_{\text{ind}}$ genomic relatedness matrix (GRM). To estimate the relatedness between individuals i and i' (i.e., entry $G_{ii'}$ of $\mathbf{G}$), we used the estimator

$$G_{ii'} = \frac{\sum_{m=1}^{M}\left(v_{im} - 2\hat{p}_m\right)\left(v_{i'm} - 2\hat{p}_m\right)}{2\sum_{m=1}^{M}\hat{p}_m\left(1-\hat{p}_m\right)}, \qquad (2)$$

as given by VanRaden (2008), where M is the number of SNPs, v_{im} is the number (0, 1, or 2) of minor alleles individual i has at SNP m, and $\hat{p}_m$ is the estimated minor allele frequency (MAF) at SNP m. Note that we always computed the allele frequencies $\hat{p}_m$ from the full population in the given scenario (Helgeland, southern, or merged), which implies that the base population for estimates of σ^2_{G} does not necessarily equal the ensemble of individuals in the combined training and test sets (Hayes et al., 2009b; Legarra, 2016). Furthermore, we ensured the positive-definiteness of $\mathbf{G}$ (e.g., Hollifield et al., 2022) by adding the lowest eigenvalue d of the given GRM to its diagonal (the exact values added ranged from 10^{-9} to 0.0066, see the tables in Appendix S6 for exact values in a given model).

Before we estimated the GRMs $\mathbf{G}$ as described above, we used `PLINK 1.9` (Chang et al., 2015) to apply quality control filters to the SNP data. Individuals were filtered for call rate (> 0.95), while SNPs were filtered for call rate (> 0.9) and MAF (> 0.01). The remaining missing variant calls were not imputed, but terms in equation 2 where either v_{im} or $v_{im'}$ was a missing call were skipped in the computation of the entries of $\mathbf{G}$. The quality control filters were always applied twice: once for the scenario overall, and again for the subset of data used in a specific GP model (see below). Thus, the exact subset of SNPs and individuals differ slightly between different models (see the tables in Appendix S6 for the exact sample sizes and numbers of SNPs in each model).

Within- and across-population genomic prediction

Generally, in statistical prediction models, we distinguish between the data used to fit the model, which we label the *training set*, and the data used to assess the model, which we label the *test set*. Assessment of the prediction accuracy on the test set thus reflects the model's performance on unseen data. How much information the training set provides about the test set is crucial in determining the performance of a prediction model, and the same is true in GP (Akdemir & Isidro y Sánchez, 2019; Pszczola et al., 2012; Rio et al., 2021). As the goal of this study is to demonstrate how across-population GP accuracy is affected by various factors, we draw extra attention to the design of our training and test sets. In each of the three scenarios, we considered two types of models: *within*-population GP and *across*-population GP. In within-population GP, the training and test samples are drawn from the same (meta-)population, whereas in across-population prediction, the training and test samples come from two distinct (meta-)populations (Figure 2).

We assessed the within-population models using 10-fold cross-validation (CV), that is, by randomly partitioning the data in a given scenario into 10 equally sized folds, and then fitting 10 GP models where in a given model, one of the folds was used as the test set, and the remaining nine were together used as the training set. Our within-population GP models thus emulated predicting the genetic values of genotyped but non-phenotyped individuals from an already-sampled (meta-)population. Aside from partitioning the folds by individuals rather than by single measurements to prevent repeatedly measured individuals from ending up in more than one fold, we ignored the presence of any population structure when creating the folds for within-population GP. Thus, sparrows from all the different islands in a given scenario appeared in both the training and test set for a given within-population model. Notably, this approach causes closely related individuals to end up in different folds, giving high relatedness between training and test set individuals, and ensures that the environments experienced by the test set individuals were also represented in the training set.

Conversely, for our across-population models, we drew the training and test individuals from different (meta-)populations, with varying designs for the different scenarios. We treated each island as a distinct population, and groups of islands as meta-populations, meaning that each scenario represents a meta-population of islands connected by some gene flow. In Scenario 1 (Helgeland), we used the following designs. For one, we performed a leave-one-island-out CV, where for a given model, all sparrows from one island made up the test set, and the sparrows from all other islands were used to train the model. We also fitted models where the farm islands together made up the test set and the

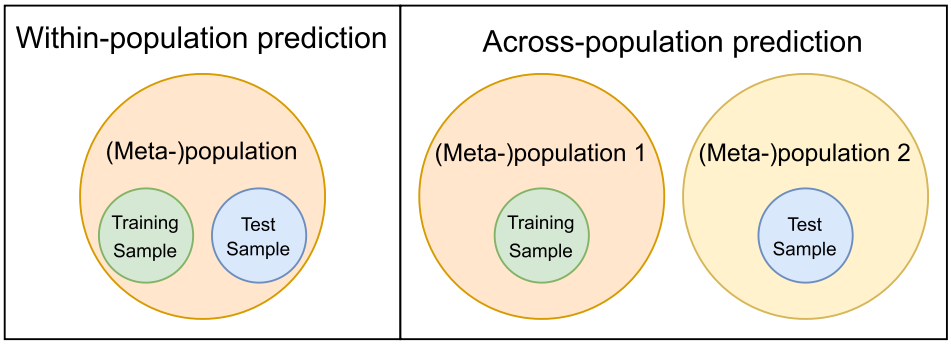

Figure 2. Conceptual diagram showing the difference between within-population and across-population GP. In within-population GP, the training and test samples are drawn from the same (meta-)population, whereas in across-population GP, the training and test samples are drawn from two different (meta-)populations.

non-farm islands together made up the training set, and vice versa. Finally, for each island in the scenario, we also used that island as a test island for models fitted on all the farm islands (except the test island if it was a farm island), and models fitted on all the non-farm islands (except the test island if it was a non-farm island). For Scenario 2 (southern islands), we again performed a leave-one-island-out CV. In Scenario 3 (merged), we predicted the genetic values in the full Helgeland meta-population based on all the sparrows from the southern islands, and vice versa. Thus, the across-population models in the three scenarios captured varying degrees of differentiation between the (meta-)populations across which we predicted, both environmentally and genetically. Note that most populations contained some level of admixture (Aase et al., 2022), due to either natural dispersal or translocation (Kvalnes et al., 2017; Nafstad et al., 2023; Ranke et al., 2020, 2021; Saatoglu et al., 2021). There was thus some ambiguity about whether to assign individuals that were recorded on multiple islands to the test or training set in a given model. We dealt with this issue by using individuals' hatch island to partition the full set of individuals into disjoint sub-populations, and letting the test set of a given model be the set of sparrows with hatch island corresponding to the test set islands. For sparrows that were observed as nestlings, the hatch island is known. Otherwise, we used the first island of observation as the sparrow's hatch island, as this is most likely to be the individual's true hatch island due to the sedentary nature and relatively low dispersal rate in house sparrows (Ranke et al., 2021; Saatoglu et al., 2021). We note that, due to dispersal, this hatch island-based assignment in rare cases resulted in training set individuals contributing phenotypes that were measured on test set islands (and vice versa), which could in the worst case cause some information leakage between the training and test sets. However, such measurements would primarily be problematic in the presence of genotype-by-environment interactions, and similar information leakage could occur, but be undetected, in studies with less detailed tracking of dispersal.

Accuracy metric

Fitting model (1) in a Bayesian framework provided estimates of the full posterior distributions for all genetic values g_i, and we used the posterior means, denoted $\hat{g}_i$, as point estimates of g_i. A common accuracy measure for GP models is

$$r = \mathrm{Corr}\,(g_i,\ \hat{g}_i)\,,$$

that is, the correlation between true and estimated genetic values in the test set of a given model. However, in real data studies the true g_i is an unknown latent variable and is not explicitly measurable. Therefore, one cannot *directly* compute r, except in simulation studies where g_i would be known. Fortunately, scaling the easily estimable correlation $\mathrm{Corr}(y_{ijk},\ \hat{g}_i)$ by an appropriate factor allows us to indirectly estimate r, given that the model is correctly specified (see Appendix S2). In the literature, the scaling factor is commonly equated to h, the square root of narrow-sense heritability of the focal trait. However, as we show in Appendix S2, the appropriate definition of the factor should instead be

$$\lambda = \sqrt{\frac{\sigma^2_{\mathrm{G}}}{\mathrm{Var}(y_{ijk})}}\,,$$

which does not necessarily equal h (since only biologically relevant sources of variation should be included in the denominator of h^2, see de Villemereuil et al., 2018; Wilson, 2008, whereas λ also contains measurement error). We therefore measured the GP accuracy r in a given test set as

$$\hat{r} = \frac{\widehat{\mathrm{Corr}}\,(y_{ijk},\ \hat{g}_i)}{\hat{\lambda}}\,, \tag{3}$$

where $\widehat{\mathrm{Corr}}\,(y_{ijk},\ \hat{g}_i)$ denotes the sample correlation between phenotypic measurements and estimated genetic values in the test set, and $\hat{\lambda} = \sqrt{\frac{\hat{\sigma}^2_{\mathrm{G}}}{\widehat{\mathrm{Var}}(y_{ijk})}}$, with $\hat{\sigma}^2_{\mathrm{G}}$ and $\widehat{\mathrm{Var}}(y_{ijk})$ being the posterior mean of σ^2_{G} and the sample variance of the phenotypic measurements, respectively. To estimate $\hat{r}$ as accurately as possible, both $\hat{\sigma}^2_{\mathrm{G}}$ and $\widehat{\mathrm{Var}}(y_{ijk})$ correspond only to individuals in the test set for a model. For the $\hat{\sigma}^2_{\mathrm{G}}$, this was achieved by fitting separate genomic animal models with the full data set of a given scenario, and then subsampling the posteriors for individuals in the test set (see Appendix S1), while $\widehat{\mathrm{Var}}(y_{ijk})$ is the sample variance of measurements in the test set. We note that using posterior means as point estimates of σ^2_{G} could introduce bias in $\hat{\lambda}$ due to Jensen's inequality, depending on the shape of the posterior of σ^2_{G}. We did not run into this issue with posterior means since our posteriors for σ^2_{G} were narrow, but in

general posterior medians and modes have better statistical properties (see He & Hodges, 2008; Pick et al., 2023).

Note that the two vectors containing the realizations of the random variables y_{ijk} and $\hat{g}_i$, that is, the phenotypic measurements and point estimates for g_i, respectively, have different lengths due to the repeated measurements of the response. Therefore, when we estimated the sample correlation $\widehat{\mathsf{Corr}}\left(y_{ijk},\ \hat{g}_i\right)$, we increased the length of the vector of realizations of $\hat{g}_i$ by appropriately repeating elements to match the length of the vector of realizations of y_{ijk}, so that the indices i correspond. However, this repetition of elements causes $\widehat{\mathsf{Corr}}\left(y_{ijk},\ \hat{g}_i\right)$ to be attenuated (biased towards zero, see, e.g., Carroll et al., 1995). Fortunately, dividing by $\hat{\lambda}$ as suggested in equation 3 properly corrects for the attenuation in the correlation estimate (see the scaling factor derivation in Appendix S2). Dividing by λ thus simultaneously provides an indirect estimator for r, and also corrects for the correlation-attenuation induced by repeated measures.

In addition to the point estimates $\hat{r}$ of the prediction accuracies, one can also calculate the uncertainty in $\hat{r}$. For the within-population models we can use the CV standard deviation, but the method for attaining the uncertainties in $\hat{r}$ for the across-population models is more involved and involves resampling from the GP model posteriors. We describe this resampling-based method in Appendix S2.

Formulas for expected accuracy

When planning or assessing a GP study, it may be useful to forecast how high the accuracy will be in a given scenario. Such theoretical a priori estimates of the GP accuracy can tell us how much data we need to collect to attain a certain level of accuracy, or whether the model is performing as well as expected. To this end, heuristic formulas that quantify $\mathsf{E}(r)$, the *expected* accuracy from a GP model, have been developed. Most such expressions are variations on the formula

$$\mathsf{E}(r) \approx \sqrt{\frac{1}{1+\frac{M_e}{Nh^2}}}, \tag{4}$$

originally presented by Daetwyler et al. (2008, 2010), where h^2 is the heritability of the trait, N is the number of individuals in the training set and M_e is the *effective number of independent chromosome segments*. The parameter M_e can thus be interpreted as the number of independent effects the model has to estimate from the data, and is positively related to effective population size and the length of the genome (Brard & Ricard, 2015). Alternatively, M_e is sometimes instead viewed as a model-specific parameter capturing how suitable the training set is for a given test set (Lee et al., 2017).

At first glance, formulas like equation 4 appear to be helpful tools in assessing our GP model accuracies (as has been done before in wild systems, see Ashraf et al., 2022). However, use of formulas for expected accuracy is in practice severely complicated by various interconnected issues that we outline below. One set of issues relates to the underlying assumptions of the formula, while another involves the interpretations of the formula parameters. Additional issues arise in across-population GP. As we will argue, these issues necessitate that the user make some arbitrary choices, which can potentially result in equation 4 failing to predict the realized GP accuracies (as demonstrated in Appendix S3). What follows is an extended discussion of why we therefore prefer not to compare our observed accuracies (3) to the expected accuracies from (4).

Issues related to formula assumptions

Equation 4 and its aforementioned generalizations impose strict assumptions of an "ideal" scenario, namely an unstructured population of unrelated individuals sampled once in a homogeneous environment (Daetwyler et al., 2008). First, the assumption that we are working with a sample of unrelated individuals is particularly unrealistic in our scenario. Relatedness between individuals within and across the test and training sets impacts GP accuracy (Dekkers et al., 2021; Habier et al., 2007; Pszczola et al., 2012), so the assumption of individuals being unrelated could bias the expected accuracy. Second, the assumption that no heterogeneous environmental effects are present in the system implies that a phenotype y_i can simply be decomposed as $y_i = g_i + \varepsilon_i$, where ε_i is an unstructured environmental residual, but in wild study systems, the situation is usually more complicated. For example, one has little opportunity to control the environment, and thus has to account for environmental confounders as additional fixed or random effects in the model (Kruuk & Hadfield, 2007). In natural populations, these environmental effects can confound estimates of genetic effects by creating structured phenotypic covariances between related individuals, and accounting for them is therefore important. In addition, the sampling design of a study might necessitate additional effects in the model, such as accounting for measurement error (Ponzi et al., 2018). In any case, it becomes unclear how to interpret the formula's parameters in the presence of structured non-genetic effects in the model (see below), and certain steps in the derivations of formula (4) become invalid. In particular, it is assumed that the relation $r^2_{y,\hat{y}} = h^2 r^2$ between *phenotypic* prediction accuracy $r_{y,\hat{y}}$ and GP accuracy r holds. However, as shown in Appendix S3, $r^2_{y,\hat{y}} = h^2 r^2$ only holds when $y_i = g_i + \varepsilon_i$, which is not true if we have other effects than g_i in the model.

Other limiting assumptions in the original derivation of (4) by Daetwyler et al. (2008) have been relaxed thanks to various generalizations (e.g., Goddard et al., 2011; Wientjes et al., 2015; Wray et al., 2013, 2019). For example, it is possible to account for the presence of relatedness by disentangling the genomic and pedigree-based sources of explanatory power (Dekkers et al., 2021). However, the respective method requires fitting additional models that rely on information from pedigrees, whereas one of the main advantages of using genomic data in wild populations is that the laborious construction of pedigrees can be circumvented. Regarding the assumption of $y_i = g_i + \varepsilon_i$, we note that it is common to pre-fit a mixed model with all non-genetic effects, and treating the identity effect from this non-genetic model as the new pseudo-phenotypes y^*_i in a GP model (e.g., Ashraf et al., 2022; Dadousis et al., 2014; de Oliveira et al., 2023; Hidalgo et al., 2016; Hunter et al., 2022). While joint modeling of genetic and environmental effects in a single step would be theoretically superior, the two-step approach is often used due to software limitations where specialized GP tools cannot accommodate the full joint model (e.g., Yin et al., 2025). By performing the pre-fitting step, the formula's assumption of $y^*_i = g^*_i + \varepsilon^*_i$ is true in the *second* model, given that we have

accounted for all non-genetic effects. However, the variance explained in the first step is not accounted for in subsequent analyses; thus, such a two-step approach could potentially bias the GP results (see, e.g., Freckleton, 2002).

Issues related to formula parameters

Here, we discuss issues related to the parameters of equation 4, particularly M_e. A range of estimators exists for M_e, each with somewhat differing assumptions (Brard & Ricard, 2015). The parameter M_e has, for instance, been estimated from effective population sizes, LD, and genome length (Daetwyler et al., 2010; Goddard et al., 2011), from the variance of relatedness (Lee et al., 2017; Wientjes et al., 2015), or from rearranging equation 4 and inserting for $\mathsf{E}(r)$ a realized $\hat{r}$ obtained in a previous model fitted to data from the same system (Dekkers et al., 2021). The problem is that it is not possible to know a priori which estimator for M_e is most appropriate in a particular scenario (Brard & Ricard, 2015 and see Appendix S3). Therefore, the only way to validate a given M_e estimate is to fit the model and compare the expected accuracy from (4) with the realized accuracy, but choosing an M_e estimator with this method would be circular reasoning. We also note that defining the heritability h^2 in the presence of additional fixed and random effects is not always straightforward (see, e.g., de Villemereuil et al., 2018; Wilson, 2008, and Appendix S3).

Expected accuracy in across-population GP

In across-population GP, there are additional issues with current expected accuracy formulas, and the standard expected accuracy formula, equation 4 breaks down in the across-population case. Wientjes et al. (2015) therefore proposed scaling equation 4 by the trait-specific genetic correlation between the two populations to correct for the decrease in accuracy when predicting across populations. However, estimating such genetic correlations is only possible if we have genotypic *and* phenotypic data from both populations, but if one has measured the phenotypes in the test population, then there is no reason to do a pure across-population prediction. In Appendix S3, we illustrate that even this scaling-approach failed at forecasting across-population GP accuracy for the house sparrow data. Furthermore, additional complications with expected accuracy formulas arise in across-population GP, such as the possibility that h^2 is not equal in the two populations, or the presence of genotype-by-environment interactions, both of which can only be detected using phenotypes from both populations. Other commonly cited reasons for the decrease in accuracy, such as population differences in LD and MAFs (Wang et al., 2020), are also not accounted for by the formula in equation 4. The complexity of these factors, and interactions between them, suggests that simple predictive formulas for across-population GP accuracy may be unattainable.

Population differentiation measures

The challenges with finding useful formulas to predict GP accuracy, especially across populations, reflect that the accuracy is impacted by underlying mechanisms that are not yet fully understood. To investigate which factors impacted across-population GP accuracies in our house sparrow applications, we estimated several measures of population differentiation between the training and test sets of each of the across-population models. As detailed below, we computed summaries of across-population genomic relatedness, a measure of population differences in LD, a measure of population structure strength based on LD inflation, and pairwise population genetic fixation indices (Wright's F_{st}, which reflect differences in allele frequencies). Similar to the aforementioned expected accuracy formulas, these measures can be computed from genomic data only (as opposed to also requiring phenotypic measurements), which allows them to be used in an a priori consideration of whether GP is worthwhile. In Appendix S4, we additionally present estimates of trait-specific genetic correlations between training and test populations.

Across-population genomic relatedness

Relatedness is expected to be a key factor in determining GP accuracy (Dekkers et al., 2021; Pszczola et al., 2012). For each model, we therefore computed summary statistics on how related individuals in the training sets are to individuals in the test set. In other words, for a given model, we considered a subset $G_{ii',\mathrm{ac}}$ of entries of the GRM defined in equation 2, such that individual i is in the training set and i' is in the test set. To summarize these relatednesses, we computed the sample mean, which we denote $\bar{G}_{\mathrm{ac}} = \widehat{\mathsf{E}}(G_{ii',\mathrm{ac}})$. We also consider another summary statistic, the sample precision $\tilde{G}_{\mathrm{ac}} = \frac{1}{\widehat{\mathsf{Var}}(G_{ii',\mathrm{ac}})}$, which has previously been considered important in determining GP accuracy (and indeed been employed as an estimator of the M_e parameter, see Lee et al., 2017; Wientjes et al., 2015).

Population differences in linkage disequilibrium

The quality of GP models fundamentally relies on SNPs tagging quantitative trait loci (QTL) for the trait of interest (de los Campos et al., 2015). Thus, a critical type of difference between populations can lie in the patterns of LD within the respective training and test populations (de Roos et al., 2008). If there is a difference between the two populations in what loci are tagged by a given SNP, or the strength of that association, then the SNP-effects will in practice differ between the populations. To measure such discrepancies in LD between training and test sets, we first look at the difference in LD for a given pair of SNPs (m, m')

$$\delta\mathrm{LD}_{m,m'} = \widehat{\mathsf{Corr}}\left(v_{im}, v_{im'}\right)^{(\mathrm{train})} - \widehat{\mathsf{Corr}}\left(v_{im}, v_{im'}\right)^{(\mathrm{test})} ,$$

where v_{im} and $v_{im'}$ are the numbers of minor alleles that individual i has at m and m', respectively. The superscripts indicate whether the Pearson correlations were computed using the individuals in the training or test set. The further the two populations have diverged, the more we expect the LD patterns in the populations to differ on average, so we consider the absolute mean difference in LD over all pairs

$$\delta\mathrm{LD} = \frac{1}{\#\mathrm{pairs}} \sum_{m,m'} \left|\delta\mathrm{LD}_{m,m'}\right| , \qquad (5)$$

where we have taken the absolute value since it does not matter which direction $\delta\mathrm{LD}_{m,m'}$ changed. To reduce the computational burden, and since we are interested in how SNPs encode for nearby QTL, we limited our estimation of δLD to SNP pairs located within 50 kb of each another (similar to Lundregan et al., 2018). Equation 5 captures how consistent the LD structure is between training and test populations, where large values of δLD suggest that SNP pairs have

different LD patterns in the two populations, which could undermine GP accuracy.

Because the presence of population structure is expected to inflate LD in a meta-population (Nei & Li, 1973), we also used LD-differences to measure the level of genetic differentiation between the training and test sets. Namely, we investigated how $\widehat{\text{Corr}}\,(v_{im}, v_{im'})$ changes depending on whether population structure is taken into account or not. We define

$$\Delta\text{LD}_{m,m'} = \widehat{\text{Corr}}\,(v_{im}, v_{im'})^{(\text{unpooled})} - \widehat{\text{Corr}}\,(v_{im}, v_{im'})^{(\text{pooled})},$$

where the first term in $\Delta\text{LD}_{m,m'}$ disregards the population structure by calculating $\widehat{\text{Corr}}\,(v_{im}, v_{im'})$ using all individuals in the combined training and test set as if they were drawn from a panmictic population, while the second term respects the distinction between training and test sets by accounting for population structure through pooling, that is, by properly aggregating the LD statistics from the training and test set. In other words, the numerator of $\widehat{\text{Corr}}\,(v_{im}, v_{im'})^{(\text{pooled})}$ is the pooled covariance

$$\widehat{\text{Cov}}\,(v_{im}, v_{im'})^{(\text{pooled})} = \frac{(N_{\text{train}} - 1)\cdot\widehat{\text{Cov}}\,(v_{im}, v_{im'})^{(\text{train})} + (N_{\text{test}} - 1)\cdot\widehat{\text{Cov}}\,(v_{im}, v_{im'})^{(\text{test})}}{N_{\text{train}} + N_{\text{test}} - 2},$$

and the standard deviations in the denominator of $\widehat{\text{Corr}}\,(v_{im}, v_{im'})^{(\text{pooled})}$ are treated similarly. As in equation 5, we combine the $\Delta\text{LD}_{m,m'}$ for different pairs of SNP using the absolute mean

$$\Delta\text{LD} = \frac{1}{\#\text{pairs}} \sum_{m,m'} \left|\Delta\text{LD}_{m,m'}\right|, \tag{6}$$

again restricting the calculation to SNP pairs within 50 kb. A high ΔLD indicates that LD was greatly inflated when not accounting for population structure, indicating large genetic differentiation between the training and test set, which in turn could impact GP accuracy negatively. To compute the correlations $\widehat{\text{Corr}}\,(v_{im}, v_{im'})$ for the various sub-populations, we used `PLINK 1.9` (Chang et al., 2015).

Pairwise population genetic fixation indices

F_{st} is a measure of genetic differentiation between populations, which essentially reflects differences in allele frequencies (Holsinger & Weir, 2009). Weir and Goudet (2017) present an expression for finding locus-specific estimates of F_{st} for pairs of populations, which we used to calculate the F_{st} between all pairs of training and test sets in our across-population GP models. We aggregated the locus-specific estimates using the method Weir and Goudet (2017) denote the "weighted" average, resulting in the F_{st} estimates

$$F_{st} = \frac{\sum_{m=1}^{M}\left(\hat{p}_{m,1} - \hat{p}_{m,2}\right)^2}{\sum_{m=1}^{M}\left[\hat{p}_{m,1}\left(1 - \hat{p}_{m,2}\right) + \hat{p}_{m,2}\left(1 - \hat{p}_{m,1}\right)\right]},$$

where $\hat{p}_{m,1}$ and $\hat{p}_{m,2}$ are the allele frequencies for SNP m in the training and test set, respectively. For each model, we estimated $\hat{p}_{m,1}$ and $\hat{p}_{m,2}$ with `PLINK 1.9` (Chang et al., 2015).

Results

Within- and across-population GP accuracy

Using the observed accuracies from each fitted GP model, as defined in equation 3, we found that within-population GP generally achieved higher accuracies than across-population GP (Figure 3). Despite the difference in training set sample sizes (see the tables in Appendix S6), the within-population accuracies for a given trait were relatively similar in the three scenarios, with small differences in the mean accuracies relative to the variation. Conversely, the accuracies across populations in the Helgeland scenario were usually lower than in the within-populations scenarios, but highly variable, while the models from the southern and merged scenarios consistently performed more poorly in spite of the comparable training set sample sizes. The within-population accuracies were similar for the three traits, as were the across-population accuracies. However, the traits differed in the variability of the accuracies, both within and across populations, as accuracies for wing length were more variable than the accuracies for body mass, and the accuracies for tarsus length were the least variable (Figure 3). We note that some realized across-population accuracies fall below zero, possibly due to sampling variance (see uncertainty measures in Appendix S2).

Factors impacting across-population accuracy

We plotted the across-populations accuracies (i.e., the same ensemble of points as in the right column of panels in Figure 3) against various measures that potentially have an effect on accuracy (Figure 4). We found a slight trend for the across-population prediction accuracy to increase with training set size N (Figure 4A). However, accuracies were highly variable around this trend. For example, the models in Scenario 3 that were trained on the full Helgeland meta-population had the highest N, but also contained cases with some of the lowest accuracies. Higher mean relatedness $\overline{G}_{\text{ac}}$ was associated with higher accuracy (Figure 4B). Notably, most of the across-population models had negative $\overline{G}_{\text{ac}}$, which can be interpreted as the average pair of across-population training and test individuals in those models sharing *fewer* alleles and thus being less related than would be expected at random. In other words, the individuals are—unsurprisingly—less related than if the training and test populations together formed a large panmictic population. The models formed two clusters of points, where the cluster with higher $\overline{G}_{\text{ac}}$ corresponded to every model that was tested on a non-farm island and trained on the remaining non-farm islands in the Helgeland system. This cluster contained across-population models that consistently yielded high prediction accuracy. Also, the sample precision $\tilde{G}_{\text{ac}}$ seemed associated with across-population accuracy, and the models with high accuracy had comparatively low $\tilde{G}_{\text{ac}}$ (Figure 4C). Note that the models from Scenario 2 (southern islands) were clustered together in Figure 4C, as were the models from Scenario 3 (merged), which had comparatively high $\tilde{G}_{\text{ac}}$ and low accuracy. Surprisingly, LD-differentiation measured by δLD did not seem to impact across-population accuracy (Figure 4D), but we did find that across-population accuracies appeared to decrease with higher ΔLD (Figure 4E). Finally, across-population accuracy seemed not to be associated with pairwise F_{st} between the training and test populations (Figure 4F). The cluster of points with higher F_{st} correspond to all the models in the Helgeland scenario that had the test island Aldra, a highly inbred farm habitat island which was recently colonized by house sparrows (Billing et al., 2012). Note that there was a wide scatter of points

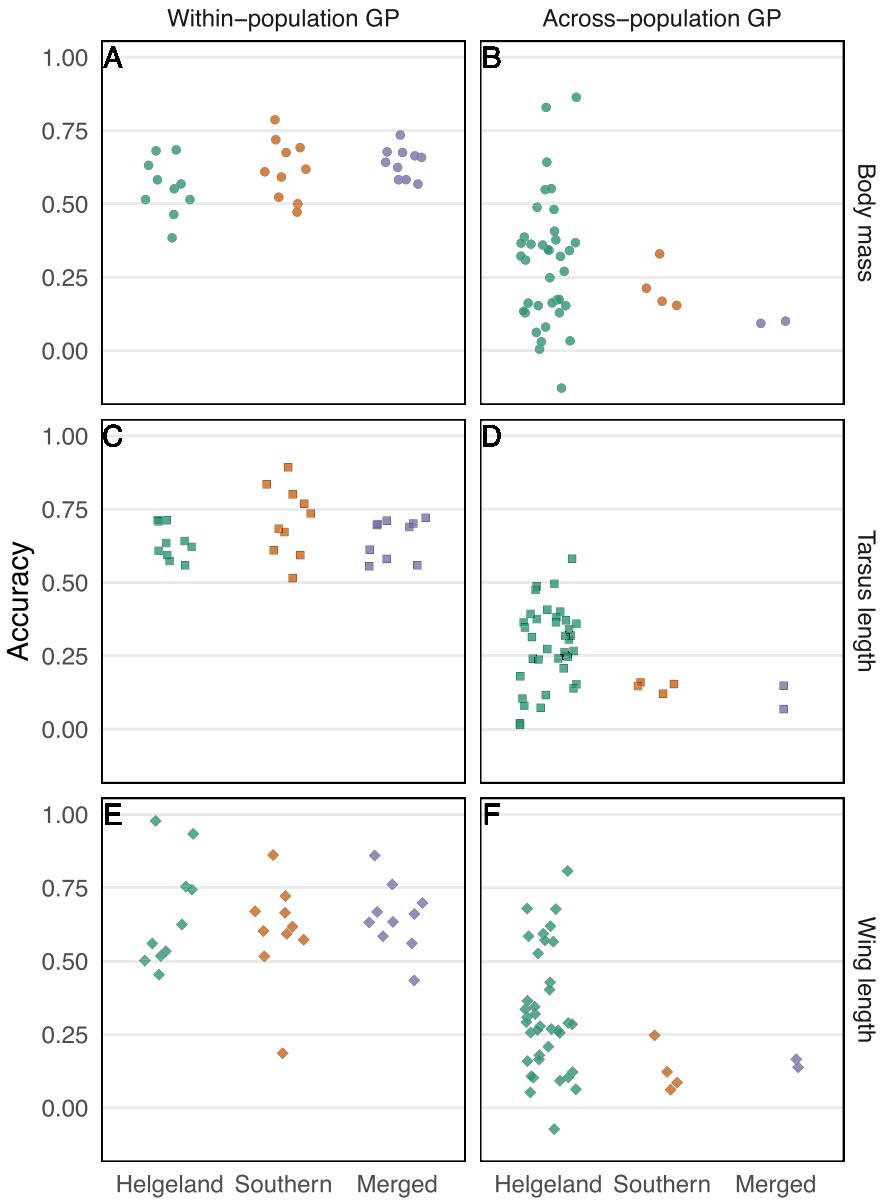

Figure 3. The accuracies from each GP model, as defined in equation 3, for different scenarios and traits in the house sparrow systems. Models in panels (A) and (B) use body mass as the response, (C) and (D) use tarsus length, and (E) and (F) use wing length. The left column of panels corresponds to within-population GP models, where each point is the accuracy for a single fold in the 10-fold CV. The right column of panels corresponds to the accuracies from the across-population GP models. Inside each panel, the points in green (left) correspond to accuracies from the Helgeland scenario, the points in orange (middle) correspond to accuracies from the southern scenario, and the points in blue (right) correspond to accuracies from the merged scenario. The points are jittered horizontally to avoid overlap.

around the smooth lines in all panels in Figure 4, indicating that none of the measures individually accounted for much variation in observed accuracy. In Appendix S4, we also investigated the impact of trait-specific genetic correlations between populations, and of heritability-related statistics, on across-population accuracy, while in Appendix S5, we performed variable importance analyses, using both multiple linear regression and GAMs, for the impact of the statistics in Figure 4, and $\mathsf{E}(r)$, on across-population accuracy. In brief, Appendix S5 shows that F_{st} and δLD are highly correlated, as are ΔLD and $\mathsf{E}(r)$, and that the parameters in Figure 4 explained less than half of the variance in across-population accuracy, but more variance than was explained by $\mathsf{E}(r)$. Notably, while δLD and F_{st} showed little apparent association with accuracy in univariate analysis (Figures 4D and F), the GAM analysis in Appendix S5 indicated that they had relatively high variable importance in a non-linear multivariate context. Appendix S6 contains detailed tables with results from all fitted GP models.

Discussion

In this article, we applied GP models to morphological phenotypes in wild populations of house sparrows. As the use of GP in wild populations is still in its infancy, our results have important implications for future applications. We demonstrated that GP achieves relatively high accuracies when applied within populations. GP is thus a promising tool for

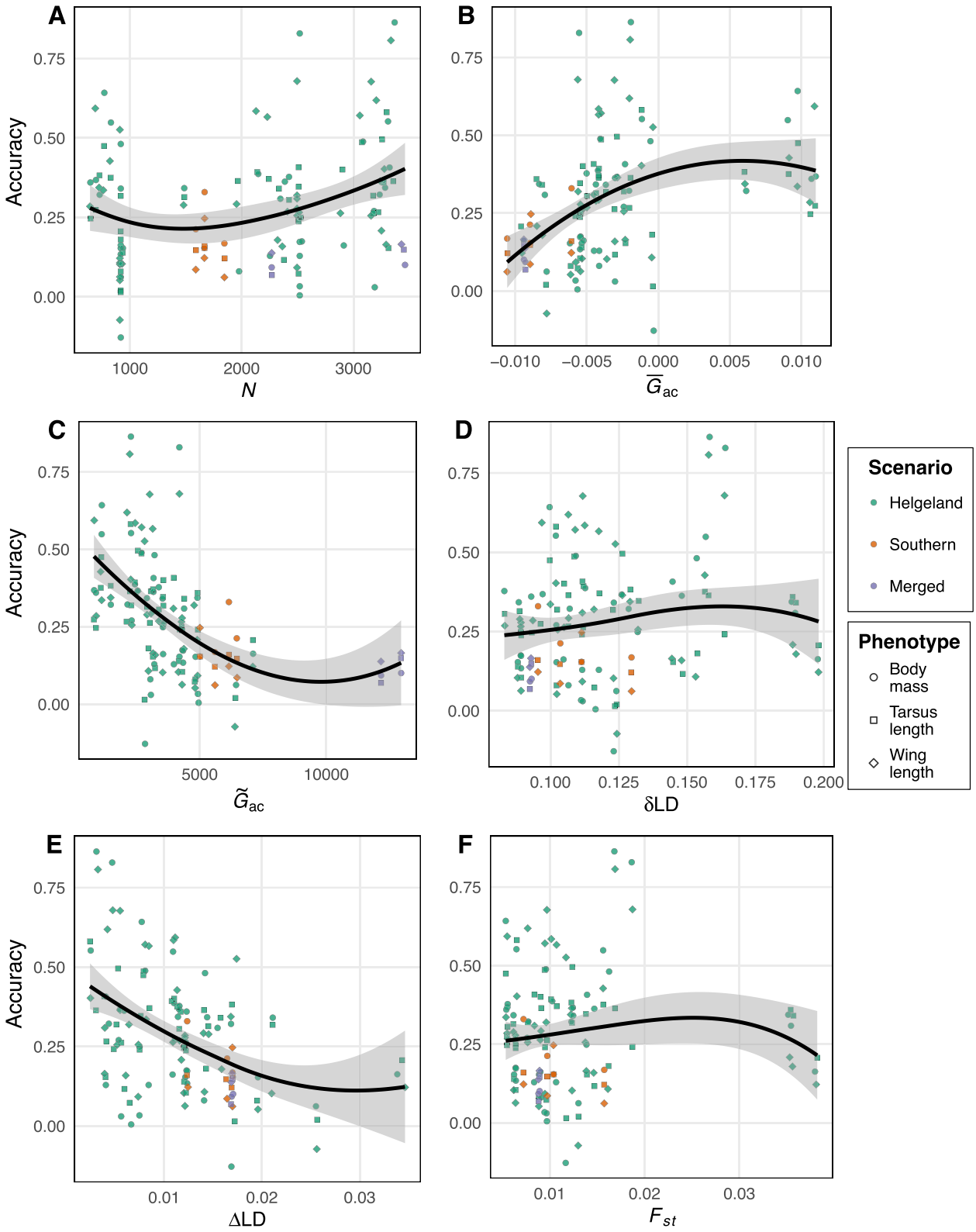

Figure 4. Scatter plots of across-population prediction accuracies for all scenarios and traits, plotted against various statistics. The statistics on the *x*-axis correspond to the number of individuals *N* in the training set (A), the mean relatedness $\overline{G}_{ac}$ between individuals in the training and test set (B), the precision $\tilde{G}_{ac}$ of relatedness between individuals in the training and test set (C), δLD, a measure of LD-differentiation between the populations (D), ΔLD, a measure of population structure between the training and test set (E), and the population-pairwise fixation index F_{st} (F). The smoothing lines are produced with local regression.

the field of evolutionary ecology, as it lets us effectively predict breeding values for non-phenotyped individuals. However, our results illustrate that across-population applications should currently be carried out with caution, as existing GP approaches are considerably less accurate when applied across populations.

Since our understanding of the exact mechanisms that determine across-population GP accuracy is limited, we empirically investigated the impact of various population differentiation statistics. Several measures appear to be important to across-population accuracy, including the mean $\overline{G}_{ac}$ and precision $\tilde{G}_{ac}$ of the relatedness between individuals in the two populations and the population structure measure ΔLD. Based on our results, we cannot recommend applying GP across populations unless $\overline{G}_{ac}$ is high, $\tilde{G}_{ac}$ is low, and/or ΔLD is low, as these situations are most consistently associated with the highest observed accuracies. However, if there is only one training and one test population, it might be unclear what acceptable values for the investigated statistics are. And even for given values of the population differentiation measures, across-population accuracy is usually highly unpredictable (i.e., the *y*-axes of Figure 4 are highly variable for most values along the *x*-axes). A major problem is that we still lack an a priori understanding of when across-population GP performs well or not, as we do not have a reliable formula to predict GP accuracy in a given scenario. A fuller and more mechanistic understanding of these underlying processes is required to reliably apply GP across populations.

Our study focused on population-differentiation statistics as predictors of across-population GP accuracy, and we did not directly investigate other potential mechanisms underlying the poor across-population performance. For example, genotype-by-environment interactions and epistatic genetic effects could in practice cause SNP effect sizes to differ between populations, leading to reduced prediction accuracy. However, theoretical work suggests that differences in LD patterns and MAFs may be more important drivers of accuracy loss than differences in effect sizes (Wang et al., 2020). Moreover, differences in additive SNP effect sizes between populations should manifest as reduced trait-specific genetic correlations between populations (Wientjes et al., 2015), a factor we did investigate and found no evidence of its importance (Appendix S4 and S5). Explicitly modeling differentiations in SNP effects, rather than their downstream effects on genetic correlations, would be challenging, but could nevertheless provide insight into the biological basis of across-population GP performance.

Given that existing GP models perform better within populations, the question arises: What should be done in situations where only limited phenotypic data are available from a population of interest, but we have access to more measurements from other, distant populations? Restricting ourselves to only using within-population GP would make populations that are small and isolated—the populations with the highest risk of extinction—especially challenging applications, as sample sizes will necessarily be small. This potentially problematic issue may be counteracted by the other consequences of populations being smaller and more isolated: Within such populations, higher levels of relatedness and LD are expected to increase GP accuracy. In any case, if one has a few phenotypes from the focal population, but also a large number of measurements from *other* populations, then the training set could in practice be "augmented" with data from these other populations. This approach could be beneficial in terms of increasing the sample size of the training set, but also risks degrading model performance due to the various factors decreasing across-population accuracy. We did not consider such training set augmentation here, and instead focused on "pure" across-population GP, where the training and test data are from entirely disjoint populations. Utilizing data from both the population of interest and other populations is a more general problem that comes down to maximizing GP accuracy by using the available data in the best possible way. This gives rise to the idea of "training set optimization" methods, where one aims to determine which individuals should be included in the training set to optimize the accuracy in a given test set (e.g., Rio et al., 2021). Investigating such optimization approaches in wild populations would be an interesting focus for a future study, in particular in the context of conservation efforts.

One potential weakness of our analysis is that we only considered relatedness-based GP models (i.e., genomic animal models, G-BLUP). These models were a natural choice since they are considered the baseline GP model against which other models are often compared (e.g., Ashraf et al., 2022; Aspheim et al., 2024). Other GP methods could potentially improve general accuracy (e.g., Bayesian approaches, see Aspheim et al., 2024; Yin et al., 2025), or better deal with the problematic aspects of across-population GP. However, accuracy gains within populations for other GP compared to G-BLUP are usually marginal for highly polygenic traits like the morphological phenotypes we investigated (see Ashraf et al., 2022). It is worth pointing out that we deliberately made no special effort to account for the models being applied across populations, which explicitly violates the assumptions of G-BLUP regarding lack of population structure (see, e.g., Wolak & Reid, 2017). Instead, we straightforwardly applied the models as one would in the within-population case, rather than accounting for, for example, population differences in allele frequencies (see, e.g., Wientjes et al., 2017). Thus, our study also investigated potential negative consequences of breaking these assumptions. One promising direction is thus the explicit modeling of population structure through genetic group approaches (Aase et al., 2022; Rio et al., 2020), which define breeding values relative to multiple reference populations rather than assuming a single ill-defined panmictic meta-population. This method could provide a principled way to prioritize the genomic regions in the training data that are most informative for a given test population. Another avenue could be to incorporate ideas from theoretical developments (do O et al., 2025; Ovaskainen et al., 2011) in multi-population quantitative genetics, which aim to resolve these fundamental conceptual issues by defining relatedness within and between contemporary subpopulations that have stemmed from an ancestral panmictic population.

By considering empirical data from a wild population, we ensured that the range of situations that we investigated were both realistic and ecologically relevant. Notably, the study covers the full spectrum of possible outcomes from across-population GP, as we notably found observed accuracies throughout the interval [0,1]. On the other hand, our approach also limited the dimensions of the problem we were able to investigate. For instance, the traits we looked at are

all continuous, highly polygenic, and have similar heritabilities (Silva et al., 2017, and see Appendix S6). Future research on across-population GP in wild populations could therefore benefit from considering a broader range of traits and incorporating simulation studies with known ground truths.

Our results highlight current limitations of across-population GP, but technological advances could improve future performance. The LD (as measured by squared correlation between markers) in the Helgeland house sparrow populations has been shown to display moderate decay, with LD $\approx$ 0.32 for SNPs within 1 kb of each other, LD $\approx$ 0.15 for SNPs within 5–10 kb, and a reduction of LD to background levels within 100 kb (Elgvin et al., 2017; Hagen et al., 2020). Given our average SNP spacing of 8–9 kb, many causal variants may be in only moderate LD with the nearest genotyped SNPs. But if sequencing costs continue to decline, the transition from SNP arrays to whole-genome sequencing will increase marker density by orders of magnitude. This may improve across-population GP accuracy by either directly typing causal loci or including markers in much tighter linkage with them, ameliorating issues of LD patterns differing between populations. Better across-population GP performance could thus be achievable in the near future, but without a clear mechanistic understanding of across-population GP accuracy, it remains uncertain whether increased marker density alone will be sufficient to solve the problem.

In conclusion, GP is an emerging tool in evolutionary ecology, with the potential to effectively predict breeding values of genotyped, but not necessarily phenotyped, individuals. While GP is reliable within single (meta-)populations, more work is needed to better understand and utilize the promise of GP across populations. In particular, we need to understand the factors that enhance or limit prediction accuracy, and whether it is possible to bypass these limitations. This would allow us to fully harness the tremendous potential of genomic data using GP, in order to understand evolutionary processes across space and time and, ultimately, help address the biodiversity crisis.

Ethical statement

The research done on house sparrows was carried out in accordance with permits from the Norwegian Food Safety Authority and the Norwegian Bird Ringing Centre at Stavanger Museum, Norway. The first author would like to note that citation of literature from the field of animal breeding is not an endorsement of the practices of the meat, dairy, or egg industries. Conversely, the first author affirms the intrinsic moral value of non-human animals, and thus opposes the unnecessary and unimaginable suffering experienced by the victims of industrial animal agriculture.

Supplementary material

Supplementary material is available online at *Evolution*.

Data availability

The data underlying this article will be available on the Dryad data repository (DOI: 10.5061/dryad.bvq83bkp1) following an embargo period of 6 months from the publication date of the article. Anyone who wishes to use our data before the embargo expires can contact H. Jensen. The code used to generate the results is available at https://github.com/kennaas/across_gp.

Author contributions

K.A., S.M., and H.J. conceived the idea for the study. K.A. performed and implemented all analyses, with support from S.M., H.J., and H.B. H.J. and H.B. provided the sparrow data. K.A. wrote the manuscript, with support from S.M. and H.J. K.A., S.M., and H.J. edited the manuscript. All authors read and approved the final manuscript.

Funding

This work was possible thanks to generous funding from the Department of Mathematical Sciences at the Norwegian University of Science and Technology (NTNU), as well as by grants from the European Research Council (grant number 101169862) and the Research Council of Norway (project numbers 274930 and 302619). This work was also partly supported by the Research Council of Norway through its Centres of Excellence funding scheme (project number 223257).

Conflict of interest

The authors have no conflicts of interest to declare.

Acknowledgments

We thank the many researchers, students, and fieldworkers who helped in collecting the empirical data on house sparrows, and laboratory technicians for assistance with laboratory analyses. Genotyping on the custom house sparrow Axiom 200K and 70K SNP arrays was carried out at CIGENE, Norwegian University of Life Sciences, Norway. The computations were performed on resources provided by the NTNU IDUN/EPIC computing cluster (Själander et al., 2019). We thank Jarrod D. Hadfield for the useful discussion that inspired the LD measures in the analysis.

References

Aase, K., Jensen, H., & Muff, S. (2022). Genomic estimation of quantitative genetic parameters in wild admixed populations. *Methods in Ecology and Evolution*, *13*, 1014–1026. https://doi.org/10.1111/2041-210X.13810

Akdemir, D., & Isidro y Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports*, *9*, 1–15. https://doi.org/10.1038/s41598-018-37186-2

Araya-Ajoy, Y. G., Niskanen, A. K., Froy, H., Ranke, P. S., Kvalnes, T., Rønning, B., Pepke, M. L., Jensen, H., Ringsby, T. H., Sæther, B. E., & Wright, J. (2021). Variation in generation time reveals density regulation as an important driver of pace of life in a bird metapopulation. *Ecology Letters*, *24*, 2077–2087. https://doi.org/10.1111/ele.13835

Ashraf, B., Hunter, D. C., Bérénos, C., Ellis, P. A., Johnston, S. E., Pilkington, J. G., Pemberton, J. M., & Slate, J. (2022). Genomic prediction in the wild: A case study in Soay sheep. *Molecular Ecology*, *31*, 6541–6555. https://doi.org/10.1111/mec.16262

Aspheim, J. C. H., Aase, K., Bolstad, G. H., Jensen, H., & Muff, S. (2024). Bayesian marker-based principal component ridge regression—A flexible multipurpose framework for quantitative genetics in wild study systems. *bioRxiv*, 2024–06.

Berens, A. J., Cooper, T. L., & Lachance, J. (2017). The genomic health of ancient hominins. *Human Biology*, *89*, 7–19. https://doi.org/10.13110/humanbiology.89.1.01

Berg, J. J., & Coop, G. (2014). A population genetic signal of polygenic adaptation. *PLoS Genetics*, *10*, e1004412. https://doi.org/10.1371/journal.pgen.1004412

Billing, A. M., Lee, A. M., Skjelseth, S., Borg, Å. A., Hale, M. C., Slate, J., Pärn, H., Ringsby, T. H., Sæther, B. E., & Jensen, H. (2012). Evidence of inbreeding depression but not inbreeding avoidance in a natural house sparrow population. *Molecular Ecology*, *21*, 1487–1499. https://doi.org/10.1111/j.1365-294X.2012.05490.x

Bosse, M., Spurgin, L. G., Laine, V. N., Cole, E. F., Firth, J. A., Gienapp, P., Gosler, A. G., McMahon, K., Poissant, J., Verhagen, I., Groenen, M. A .M., van Oers, K., Sheldon, B. C., Visser, M. E., & Slate, J., (2017). Recent natural selection causes adaptive evolution of an avian polygenic trait. *Science*, *358*, 365–368.

Brard, S., & Ricard, A. (2015). Is the use of formulae a reliable way to predict the accuracy of genomic selection? *Journal of Animal Breeding and Genetics*, *132*, 207–217. https://doi.org/10.1111/jbg.12123

Burnett, H. A., et al. (2025). Addressing cross-platform genotype data compatibility and sample mix-ups to assemble a large metapopulation scale SNP dataset and pedigree in a wild vertebrate. in preparation.

Carlson, M. O., Rice, D. P., Berg, J. J., & Steinrücken, M. (2022). Polygenic score accuracy in ancient samples: Quantifying the effects of allelic turnover. *PLoS Genetics*, *18*, e1010170. https://doi.org/10.1371/journal.pgen.1010170

Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement error in nonlinear models*. (*105*). CRC Press.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*, s13742–015. https://doi.org/10.1186/s13742-015-0047-8

Charmantier, A., Garant, D., & Kruuk, L. E. B. (2014). *Quantitative genetics in the wild*. Oxford University Press.

Choi, S. W., Mak, T. S. H., & O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nature Protocols*, *15*, 2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Cox, S. L., Ruff, C. B., Maier, R. M., & Mathieson, I. (2019). Genetic contributions to variation in human stature in prehistoric Europe. *Proceedings of the National Academy of Sciences*, *116*, 21484–21492. https://doi.org/10.1073/pnas.1910606116

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*, 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Dadousis, C., Veerkamp, R. F., Heringstad, B., Pszczola, M., & Calus, M. P. L. (2014). A comparison of principal component regression and genomic REML for genomic prediction across populations. *Genetics Selection Evolution*, *46*, 1–14. https://doi.org/10.1186/1297-9686-46-1

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS One*, *3*, e3395. https://doi.org/10.1371/journal.pone.0003395

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, *185*, 1021–1031. https://doi.org/10.1534/genetics.110.116855

de los Campos, G., Sorensen, D., & Gianola, D. (2015). Genomic heritability: What is it? *PLoS Genetics*, *11*, e1005048. https://doi.org/10.1371/journal.pgen.1005048

de Oliveira, L. F., Brito, L. F., Marques, D. B. D., da Silva, D. A., Lopes, P. S., Dos Santos, C. G., Johnson, J. S., & Veroneze, R. (2023). Investigating the impact of non-additive genetic effects in the estimation of variance components and genomic predictions for heat tolerance and performance traits in crossbred and purebred pig populations. *BMC Genomic Data*, *24*, 76. https://doi.org/10.1186/s12863-023-01174-x

de Roos, A. P. W., Hayes, B. J., Spelman, R. J., & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics*, *179*, 1503–1512. https://doi.org/10.1534/genetics.107.084301

de Villemereuil, P., Morrissey, M. B., Nakagawa, S., & Schielzeth, H. (2018). Fixed-effect variance and the estimation of repeatabilities and heritabilities: Issues and solutions. *Journal of Evolutionary Biology*, *31*, 621–632. https://doi.org/10.1111/jeb.13232

Dekkers, J. C. M., Su, H., & Cheng, J. (2021). Predicting the accuracy of genomic predictions. *Genetics Selection Evolution*, *53*, 55. https://doi.org/10.1186/s12711-021-00647-w

Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B. J., Olde Loohuis, L. M., & Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, *618*, 774–781. https://doi.org/10.1038/s41586-023-06079-4,

do O, I., Gaggiotti, O., de Villemereuil, P., & Goudet, J. (2025). A method for identifying spatially divergent selection in structured populations. *PLoS Genetics*, *21*, e1011871. https://doi.org/10.1371/journal.pgen.1011871

Elgvin, T. O., Trier, C. N., Tørresen, O. K., Hagen, I. J., Lien, S. r., Nederbragt, A. J., Ravinet, M., Jensen, H., & Sætre, G. P. (2017). The genomic mosaicism of hybrid speciation. *Science Advances*, *3*, e1602996. https://doi.org/10.1126/sciadv.1602996

Freckleton, R. P. (2002). On the misuse of residuals in ecology: Regression of residuals vs. multiple regression. *Journal of Animal Ecology*, *71*, 542–545.

Gauzere, J., Pemberton, J. M., Slate, J., Morris, A., Morris, S., Walling, C. A., & Johnston, S. E. (2023). A polygenic basis for birth weight in a wild population of red deer (*Cervus elaphus*). *G3: Genes, Genomes, Genetics*, *13*, jkad018. https://doi.org/10.1093/g3journal/jkad018

Gienapp, P., Fior, S., Guillaume, F., Lasky, J. R., Sork, V. L., & Csilléry, K. (2017). Genomic quantitative genetics to study evolution in the wild. *Trends in Ecology and Evolution*, *32*, 897–908. https://doi.org/10.1016/j.tree.2017.09.004

Gienapp, P., Calus, M. P. L., Laine, V. N., & Visser, M. E. (2019). Genomic selection on breeding time in a wild bird population. *Evolution Letters*, *3*, 142–151. https://doi.org/10.1002/evl3.103

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, *128*, 409–421. https://doi.org/10.1111/j.1439-0388.2011.00964.x

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*, 2389–2397. https://doi.org/10.1534/genetics.107.081190

Hadfield, J. D., Wilson, A. J., Garant, D., Sheldon, B. C., & Kruuk, L. E. B. (2010). The misuse of BLUP in ecology and evolution. *The American Naturalist*, *175*, 116–125. https://doi.org/10.1086/648604

Hagen, I. J., Lien, S. R., Billing, A. M., Elgvin, T. O., Trier, C. N., Niskanen, A. K., Tarka, M., Slate, J., Sætre, G. P., & Jensen, H. (2020). A genome-wide linkage map for the house sparrow (*Passer domesticus*) provides insights into the evolutionary history of the avian genome. *Molecular Ecology Resources*, *20*, 544–559. https://doi.org/10.1111/1755-0998.13134

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., & Goddard, M. E. (2009a). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution*, *41*, 1–9. https://doi.org/10.1186/1297-9686-41-1

Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, *91*, 47–60. https://doi.org/10.1017/S0016672308009981

He, Y., & Hodges, J. S. (2008). Point estimates for variance-structure parameters in Bayesian analysis of hierarchical models. *Computa-*

tional Statistics and Data Analysis, *52*, 2560–2577. https://doi.org/10.1016/j.csda.2007.08.021

Henderson, C. R. (1984). *Applications of linear models in animal breeding*. University of Guelph Press.

Hickey, J. M., Chiurugwi, T., Mackay, I., & Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics*, *49*, 1297–1303. https://doi.org/10.1038/ng.3920

Hidalgo, A. M., Bastiaansen, J. W. M., Lopes, M. S., Calus, M. P. L., & De Koning, D. J. (2016). Accuracy of genomic prediction of purebreds for cross bred performance in pigs. *Journal of Animal Breeding and Genetics*, *133*, 443–451. https://doi.org/10.1111/jbg.12214

Hill, W. G. (2014). Applications of population genetics to animal breeding, from Wright, Fisher and Lush to genomic prediction. *Genetics*, *196*, 1–16. https://doi.org/10.1534/genetics.112.147850

Holand, A. M., Jensen, H., Tufto, J., & Moe, R. (2011). Does selection or genetic drift explain geographic differentiation of morphological characters in house sparrows *Passer domesticus*? *Genetics Research*, *93*, 367–379. https://doi.org/10.1017/S0016672311000267

Hollifield, M. K., Bermann, M., Lourenco, D., & Misztal, I. (2022). Impact of blending the genomic relationship matrix with different levels of pedigree relationships or the identity matrix on genetic evaluations. *JDS Communications*, *3*, 343–347.

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: Defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, *10*, 639–650. https://doi.org/10.1038/nrg2611

Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G. M., Buyske, S., Conti, D. V., Darst, B. F., Fornage, M., Gignoux, C., Guo, X., Haiman, C., Kenny, E. E., Kim, M., Kooperberg, C., Lange, L., Manichaikul, A., & Pasaniuc, B. (2023). Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nature Genetics*, *55*, 549–558. https://doi.org/10.1038/s41588-023-01338-6,

Hunter, D. C., Ashraf, B., Bérénos, C., Ellis, P. A., Johnston, S. E., Wilson, A. J., Pilkington, J. G., Pemberton, J. M., & Slate, J. (2022). Using genomic prediction to detect microevolutionary change of a quantitative trait. *Proceedings of the Royal Society B*, *289*, 20220330.

Irving-Pease, E. K., Muktupavela, R., Dannemann, M., & Racimo, F. (2021). Quantitative human paleogenetics: What can ancient DNA tell us about complex trait evolution? *Frontiers in Genetics*, *12*, 703541. https://doi.org/10.3389/fgene.2021.703541

Jensen, H., Moe, R., Hagen, I. J., Holand, A. M., Kekkonen, J., Tufto, J., & Sæther, B. E. (2013). Genetic variation and structure of house sparrow populations: Is there an island effect? *Molecular Ecology*, *22*, 1792–1805. https://doi.org/10.1111/mec.12226

Jensen, H., Szulkin, M., & Slate, J. (2014). Molecular quantitative genetics. (pp. 209–227), Vol. 1 of Charmantier et al. (2014).

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*, 1219–1224. https://doi.org/10.1038/s41588-018-0183-z

Kruuk, L. E. B. (2004). Estimating genetic parameters in natural populations using the "animal model". *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*, 873–890. https://doi.org/10.1098/rstb.2003.1437

Kruuk, L. E. B., & Hadfield, J. D. (2007). How to separate genetic and environmental causes of similarity between relatives. *Journal of Evolutionary Biology*, *20*, 1890–1903. https://doi.org/10.1111/j.1420-9101.2007.01377.x

Kruuk, L. E. B., Slate, J., & Wilson, A. J. (2008). New answers for old questions: The evolutionary quantitative genetics of wild animal populations. *Annual Review of Ecology, Evolution, and Systematics*, *39*, 525–548. https://doi.org/10.1146/annurev.ecolsys.39.110707.173542

Kvalnes, T., Ringsby, T. H., Jensen, H., Hagen, I. J., Rønning, B., Pärn, H., Holand, H., Engen, S., & Sæther, B. E. (2017). Reversal of response to artificial selection on body size in a wild passerine. *Evolution*, *71*, 2062–2079. https://doi.org/10.1111/evo.13277

Lee, S. H., Clark, S., & van der Werf, J. H. J. (2017). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. *PloS One*, *12*, e0189775. https://doi.org/10.1371/journal.pone.0189775

Legarra, A. (2016). Comparing estimates of genetic variance across different relationship models. *Theoretical Population Biology*, *107*, 26–30. https://doi.org/10.1016/j.tpb.2015.08.005

Lindner, M., Ramakers, J. J. C., Verhagen, I., Tomotani, B. M., Mateman, A. C., Gienapp, P., & Visser, M. E. (2023). Genotypes selected for early and late avian lay date differ in their phenotype, but not fitness, in the wild. *Science Advances*, *9*, eade6350.

Lindner, M., Verhagen, I., Mateman, A. C., van Oers, K., Laine, V. N., & Visser, M. E. (2024). Genetic and epigenetic differentiation in response to genomic selection for avian lay date. *Evolutionary Applications*, *17*, e13703. https://doi.org/10.1111/eva.13703

Lundregan, S. L., Hagen, I. J., Gohli, J., Niskanen, A. K., Kemppainen, P., Ringsby, T. H., Kvalnes, T., Pärn, H., Rønning, B., Holand, H., Ranke, P. S., Båtnes, A. S., Selvik, L. K., Lien, S., Saether, B. E., Husby, A., & Jensen, H. (2018). Inferences of genetic architecture of bill morphology in house sparrow using a high-density SNP array point to a polygenic basis. *Molecular Ecology*, *27*, 3498–3514. https://doi.org/10.1111/mec.14811

Lupi, A. S., Vazquez, A. I., & de los Campos, G. (2024). Mapping the relative accuracy of cross-ancestry prediction. *Nature Communications*, *15*, 1–14.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. vol. *1*, Sinauer Associates Incorporated.

McGaugh, S. E., Lorenz, A. J., & Flagel, L. E. (2021). The utility of genomic prediction models in evolutionary genetics. *Proceedings of the Royal Society B*, *288*, 20210693.

Marciniak, S., Bergey, C. M., Silva, A. M., Hałuszko, A., Furmanek, M., Veselka, B., Velemínskỳ, P., Vercellotti, G., Wahl, J., Zariņa, G., Longhi, C, Kolář , J., Garrido-Pena, R., Flores-Fernández, R., Herrero-Corral, A. M., Simalcsik, A., Müller, W., Sheridan, A., Miliauskienė, Ž., … Perry, G. H.. (2022). An integrative skeletal and paleogenomic analysis of stature variation suggests relatively reduced health for early European farmers. *Proceedings of the National Academy of Sciences*, *119*, e2106743119.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, *100*, 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *51*, 584–591. https://doi.org/10.1038/s41588-019-0379-x,

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2016). Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers*, *6*, 6–14. https://doi.org/10.2527/af.2016-0002

Muff, S., Niskanen, A. K., Saatoglu, D., Keller, L. F., & Jensen, H. (2019). Animal models with group-specific additive genetic variances: Extending genetic group models. *Genetics Selection Evolution*, *51*, 7. https://doi.org/10.1186/s12711-019-0449-7

Nafstad, Å. M., Rønning, B., Aase, K., Ringsby, T. H., Hagen, I. J., Ranke, P. S., Kvalnes, T., Stawski, C., Räsänen, K., Sæther, B. E., Muff, S., & Jensen, H. (2023). Spatial variation in the evolutionary potential and constraints of basal metabolic rate and body mass in a wild bird. *Journal of Evolutionary Biology*, *36*, 650–662. https://doi.org/10.1111/jeb.14164

Nei, M., & Li, W. H. (1973). Linkage disequilibrium in subdivided populations. *Genetics*, *75*, 213–219. https://doi.org/10.1093/genetics/75.1.213

Niskanen, A. K., Billing, A. M., Holand, H., Hagen, I. J., Araya-Ajoy, Y. G., Husby, A., Rønning, B., Myhre, A. M., Ranke, P. S., Kvalnes, T., Pärn, H., Ringsby, T. H., Lien, S., Sæther, B. E., Muff, S., & Jensen, H. (2020). Consistent scaling of inbreeding depression in space and time in a house sparrow metapopulation. *Proceedings of the National Academy of Sciences*, *117*, 14584–14592. https://doi.org/10.1073/pnas.1909599117

Novembre, J., & Barton, N. H. (2018). Tread lightly interpreting polygenic tests of selection. *Genetics*, *208*, 1351–1355. https://doi.org/10.1534/genetics.118.300786

Ovaskainen, O., Karhunen, M., Zheng, C., Arias, J. M. C., & Merilä, J. (2011). A new method to uncover signatures of divergent and stabilizing selection in quantitative traits. *Genetics*, *189*, 621–632. https://doi.org/10.1534/genetics.111.129387

Pick, J. L., Kasper, C., Allegue, H., Dingemanse, N. J., Dochtermann, N. A., Laskowski, K. L., Lima, M. R., Schielzeth, H., Westneat, D. F., Wright, J., & Araya-Ajoy, Y. G. (2023). Describing posterior distributions of variance components: Problems and the use of null distributions to aid interpretation. *Methods in Ecology and Evolution*, *14*, 2557–2574. https://doi.org/10.1111/2041-210X.14200

Ponzi, E., Keller, L. F., Bonnet, T., & Muff, S. (2018). Heritability, selection, and the response to selection in the presence of phenotypic measurement error: Effects, cures, and the role of repeated measurements. *Evolution*, *72*, 1992–2004. https://doi.org/10.1111/evo.13573

Pszczola, M., Strabel, T., Mulder, H. A., & Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science*, *95*, 389–400. https://doi.org/10.3168/jds.2011-4338

Ranke, P. S., Skjelseth, S., Hagen, I. J., Billing, A. M., Pedersen, A. A. B., Pärn, H., Ringsby, T. H., Sæther, B. E., & Jensen, H. (2020). Multi-generational genetic consequences of reinforcement in a bird metapopulation. *Conservation Genetics*, *21*, 603–612. https://doi.org/10.1007/s10592-020-01273-7

Ranke, P. S., Araya-Ajoy, Y. G., Ringsby, T. H., Pärn, H., Rønning, B., Jensen, H., Wright, J., & Sæther, B. E. (2021). Spatial structure and dispersal dynamics in a house sparrow metapopulation. *Journal of Animal Ecology*, *90*, 2767–2781. https://doi.org/10.1111/1365-2656.13580

Ranke, P. S., Pepke, M. L., Søraker, J. r. S., David, G., Araya-Ajoy, Y. G., Wright, J., Nafstad, Å. M., Rønning, B., Pärn, H., Ringsby, T. H., Jensen, H., & Sæther, B. E. (2024). Long-distance dispersal in the short-distance dispersing house sparrow (*Passer domesticus*). *Ecology and Evolution*, *14*, e11356. https://doi.org/10.1002/ece3.11356

Rio, S., Moreau, L., Charcosset, A., & Mary-Huard, T. (2020). Accounting for group-specific allele effects and admixture in genomic predictions: Theory and experimental evaluation in maize. *Genetics*, *216*, 27–41. https://doi.org/10.1534/genetics.120.303278

Rio, S., Gallego-Sánchez, L., Montilla-Bascón, G., Canales, F. J., Isidro y Sánchez, J., & Prats, E. (2021). Genomic prediction and training set optimization in a structured mediterranean oat population. *Theoretical and Applied Genetics*, *134*, 3595–3609. https://doi.org/10.1007/s00122-021-03916-w

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*, 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Saatoglu, D., Niskanen, A. K., Kuismin, M. O., Ranke, P. S., Hagen, I. J., Araya-Ajoy, Y. G., Kvalnes, T., Pärn, H., Rønning, B., Ringsby, T. H., Saether, B. E., Husby, A, Sillanpää, M. J., & Jensen, H. (2021). Dispersal in a house sparrow metapopulation: An integrative case study of genetic assignment calibrated with ecological data and pedigree information. *Molecular Ecology*, *30*, 4740–4756. https://doi.org/10.1111/mec.16083

Saatoglu, D., Lundregan, S. L., Fetterplace, E., Goedert, D., Husby, A., Niskanen, A. K., Muff, S., & Jensen, H. (2024). The genetic basis of dispersal in a vertebrate metapopulation. *Molecular Ecology*, *33*, e17295. https://doi.org/10.1111/mec.17295

Sæther, B. E., Ringsby, T. H., Bakke, Ø., & Solberg, E. J. (1999). Spatial and temporal variation in demography of a house sparrow metapopulation. *Journal of Animal Ecology*, *68*, 628–637. https://doi.org/10.1046/j.1365-2656.1999.00314.x

Sauve, D., Hudecki, J., Steiner, J., Wheeler, H., Lynch, C., & Chabot, A. A. (2022). Improving species conservation plans under IUCN's One Plan Approach using quantitative genetic methods. *Peer Community Journal*, *2*, e50.

Silva, C. N. S., McFarlane, S. E., Hagen, I. J., Rönnegård, L., Billing, A. M., Kvalnes, T., Kemppainen, P., Rønning, B., Ringsby, T. H., Sæther, B. E., Qvarnström, A., Ellegren, H., Jensen, H., & Husby, A. (2017). Insights into the genetic architecture of morphological traits in two passerine bird species. *Heredity*, *119*, 197–205. https://doi.org/10.1038/hdy.2017.29

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*, 1–28. https://doi.org/10.1214/16-STS576

Själander, M., Jahre, M., Tufte, G., & Reissmann, N. (2019). EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure.

Strickland, K., Matthews, B., Jónsson, Z. O., Kristjánsson, B. K., Phillips, J. S., Einarsson, Á., & Räsänen, K. (2024). Microevolutionary change in wild stickleback: Using integrative time-series data to infer responses to selection. *Proceedings of the National Academy of Sciences*, *121*, e2410324121. https://doi.org/10.1073/pnas.2410324121

Vahedi, S. M., Salek Ardetani, S., Brito, L. F., Karimi, K., Pahlavan Afshari, K., & Banabazi, M. H. (2023). Expanding the application of haplotype-based genomic predictions to the wild: A case of antibody response against *Teladorsagia circumcincta* in Soay sheep. *BMC Genomics*, *24*, 335. https://doi.org/10.1186/s12864-023-09407-0

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Verhagen, I., Gienapp, P., Laine, V. N., van Grevenhof, E. M., Mateman, A. C., van Oers, K., & Visser, M. E. (2019). Genetic and phenotypic responses to genomic selection for timing of breeding in a wild songbird. *Functional Ecology*, *33*, 1708–1721. https://doi.org/10.1111/1365-2435.13360

Walsh, B., & Lynch, M. (2018). *Evolution and selection of quantitative traits*. Oxford University Press.

Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., & Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications*, *11*, 3865. https://doi.org/10.1038/s41467-020-17719-y,

Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, *206*, 2085–2103. https://doi.org/10.1534/genetics.116.198424

Wientjes, Y. C. J., Veerkamp, R. F., Bijma, P., Bovenhuis, H., Schrooten, C., & Calus, M. P. L. (2015). Empirical and deterministic accuracies of across-population genomic prediction. *Genetics Selection Evolution*, *47*, 1–14. https://doi.org/10.1186/s12711-014-0081-5

Wientjes, Y. C. J., Bijma, P., Vandenplas, J., & Calus, M. P. L. (2017). Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics*, *207*, 503–515. https://doi.org/10.1534/genetics.117.300152

Wilson, A. J. (2008). Why h^2 does not always equal V_A/V_P? *Journal of Evolutionary Biology*, *21*, 647–650. https://doi.org/10.1111/j.1420-9101.2008.01500.x

Wolak, M. E., & Reid, J. M. (2017). Accounting for genetic differences among unknown parents in microevolutionary studies: How to include genetic groups in quantitative genetic animal models. *Journal*

of Animal Ecology, *86*, 7–20. https://doi.org/10.1111/1365-2656.12597

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, *14*, 507–515. https://doi.org/10.1038/nrg3457

Wray, N. R., Kemper, K. E., Hayes, B. J., Goddard, M. E., & Visscher, P. M. (2019). Complex trait prediction from genome data: Contrasting EBV in livestock to PRS in humans: Genomic prediction. *Genetics*, *211*, 1131–1141. https://doi.org/10.1534/genetics.119.301859

Yin, L., Zhang, H., Li, X., Zhao, S., & Liu, X. (2025). hibayes: An R package to fit individual-level, summary-level and single-step bayesian regression models for genomic prediction and genome-wide association studies. *Journal of Statistical Software*, *114*, 1–37.

Zhong, S., Dekkers, J. C. M., Fernando, R. L., & Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics*, *182*, 355–364. https://doi.org/10.1534/genetics.108.098277

Received June 17, 2025; revisions received September 22, 2025; accepted September 29, 2025
Associate Editor: Henrique Teotonio; Handling Editor: Jason Wolf