Optimising sampling design and sequencing strategy for the genomic analysis of quantitative traits in natural populations

Jefferson F. Paril[1], David J. Balding[1,2,3], Alexandre Fournier-Level[1,2]

[1]School of Biosciences, [2]Melbourne Integrative Genomics, and [3]School of Mathematics and Statistics, The University of Melbourne, Parkville 3010, Australia

Author for correspondence: Alexandre Fournier-Level (afournier@unimelb.edu.au)

Keywords: genome-wide association, individual sequencing, landscape genomics, pool sequencing, genomic prediction, simulation

Abstract

Mapping the genes underlying ecologically-relevant traits in natural populations is fundamental to develop a molecular understanding of species adaptation. Current sequencing technologies enable the characterisation of a species' genetic diversity across the landscape or even over its whole range. The relevant capture of the genetic diversity across the landscape is critical for a successful genetic mapping of traits and there are no clear guidelines on how to achieve an optimal sampling and which sequencing strategy to implement. Here we determine through simulation, the sampling scheme that maximises the power to map the genetic basis of a complex trait in an outbreeding species across an idealised landscape and draw genomic predictions for the trait, comparing individual and pool sequencing strategies. Our results show that QTL detection power and prediction accuracy are higher when more populations over the landscape are sampled and this is more cost-effectively done with pool sequencing than with individual sequencing. Additionally, we recommend sampling populations from areas of high genetic diversity. As progress in sequencing enables the integration of trait-based functional ecology into landscape genomics studies, these findings will guide study designs allowing direct measures of genetic effects in natural populations across the environment.

Introduction

Understanding how the molecular variation within species supports the evolution of functional traits is a central goal in ecology through the determination of so-called genotype-to-phenotype map. Genomic information for a species can then be leveraged to understand and eventually predict population fitness under a range of eco-evolutionary scenarios. Unfortunately, this genotype-to-phenotype map is only available for a handful of traits, and primarily in model organisms. With the improved accessibility of sequencing technologies, genome-wide association studies (GWAS) and genomic prediction (GP) are becoming straightforward approaches to understand and predict complex traits (Gondro et al., 2013). We thus coined the term GPAS (Genomic Prediction and Association Studies) to denote genome-wide association studies designed to both identify quantitative trait loci (QTL) and predict traits from genomic data. These studies rely on cost-effective, high-throughput sequencing and share the same well-established linear modelling framework. However, how to sample natural populations to train accurate GPAS models that are representative of the genetic diversity of a species is far from obvious. More insights are needed to develop an optimal strategy and move away from *ad hoc* field sampling.

Research in ecological genomics deals with the challenge of characterising the genetic basis of traits across multiple natural populations. This requires collecting sufficient genotype and experimental phenotype data to represent the species' diversity, which in practice is often performed with limited resources. This raises the problem of how to sample across a landscape to capture representative genetic variation while constrained by the total sequencing depth

attainable for a given budget. Here, we address the question: how do we allocate a fixed sequencing capacity so that the genetic information captured over the landscape leads to an optimal GPAS performance?

It is becoming easier to genotype genome-wide markers for large numbers of individuals, either through whole-genome sequencing or complexity reduction approaches such as restriction site-associated DNA sequencing (RADseq) (Baird et al., 2008) in the case of large and complex genomes. Increasing the density and number of markers for GPAS has the potential to increase QTL detection power and prediction accuracy (de Roos et al., 2009; Long & Langley, 1999). Despite the declining costs of sequencing, genotyping every individual of every population across a landscape is usually not feasible, and phenotyping remains resource-consuming. As a cost-effective alternative to sequencing individuals (Indi-seq), sequencing pools of individuals (Pool-seq) (Schlötterer et al., 2014) has gained popularity in ecology (Bastide et al., 2013; Cheng et al., 2012; Nielsen et al., 2018), evolution (Boitard et al., 2012; Fournier-Level et al., 2019; Fracassetti et al., 2015), and breeding (Beissinger et al., 2014; Bélanger et al., 2016) supported by developments in quantitative genetics (Fournier-Level et al., 2017; Guo et al., 2018; Knight, Saccone et al., 2009; Macgregor et al., 2006; Micheletti & Narum, 2018; Jinliang Yang et al., 2015).

Indi-seq generates high-resolution genomic data of a population; while Pool-seq yields low-resolution data in favour of cost reduction. Indi-seq yields individual allele information after variant calling, while Pool-seq generates allele frequency estimates for a group of individuals. Identifying when best to use one over the other is important. Pool-seq was shown to be at least

as accurate as Indi-seq in estimating genome-wide allele frequencies (Fracassetti et al., 2015; Gautier et al., 2013; Rellstab et al., 2013; Zhu et al., 2012), but it is also prone to biases in genome representation (i.e. unequal representation of different genomic regions) when sample size and depth of coverage are low (i.e. <40 individuals per pool and <50X depth) (Cutler & Jensen, 2010; Schlötterer et al., 2014). Pool-seq also loses haplotype and linkage disequilibrium (LD) information (Fariello et al., 2017) which limits the number of quantitative and population genetics models that can be used and requires the design of novel analysis methods (Cutler & Jensen, 2010). Pool-seq is more cost-effective than Indi-seq since it requires less sequencing effort to generate the same genome-wide allele frequency data (Futschik & Schlötterer, 2010; Gautier et al., 2013), particularly for non-model organisms where individuals cannot be maintained indefinitely and used in multiple experiments. Additionally, Pool-seq can include more individuals, grouped into one or a few pools and sequenced at a high depth.

There is no research on the optimal sampling strategy across a landscape for GPAS, comparing Indi-seq and Pool-seq. Quantitative genetics studies have established that large sample sizes and diverse mapping and training populations improve the power of GWAS (Visscher et al., 2012; Visscher et al., 2017; Gurdasani et al., 2019; Wojcik et al., 2019) and the accuracy of GP (Asoro et al., 2011; Bernal-Vasquez et al., 2014; Bustos-Korts et al., 2016; Rincent et al., 2017; Akdemir & Isidro-Sánchez, 2019; Edwards et al., 2019). However, the spatial component of sampling across the landscape is lacking in these studies. In landscape genomics, in addition to the recommendation to sample as many populations as possible (Santos & Gaiotto, 2020), the spatial extent of the species' dispersal should be accounted for (Riginos et al., 2016) and stratified sampling is commonly performed to represent the different environmental clines (Hoel,

1943; Li et al., 2017; Williams & Brown, 2019; Pais et al., 2020). Landscape genomics associates genotype with the environment (Li et al., 2017); while GPAS associates genotype with phenotype (Gondro et al. 2013). It follows that in GPAS, sampling every environmental cline may not be needed as long as the genetic diversity is well-represented in the samples. Additionally, landscapes and populations with low differentiation or structure allow for less samples to represent the total diversity sufficiently. Finally, it is not known whether the cost-effectiveness of Pool-seq, being the preferred genotyping approach in landscape genomics over Indi-seq (Santos & Gaiotto, 2020), retains the same cost-effectiveness in landscape-wide GPAS.

Field researchers often adapt techniques initially developed for model organisms or crops in highly controlled environments; however, with natural populations having evolved in natural environments, devising the optimal sampling strategy becomes non-trivial. Individuals and pools can be sampled from one to a few populations or from a large number of populations. Identifying which populations warrant higher resolution (i.e. individual-resolution genotype data with Indi-seq instead of group-resolution genotype data with Pool-seq), requires some prior knowledge of the spatial distribution of genetic variability across the landscape. To address this question, we simulated landscapes under different trait architectures and population genetics scenarios with the aim of providing recommendations on the optimal sampling strategies. Specifically, we aim to answer the following three questions. How many populations do we need to sample to yield optimal GPAS performance? Under which landscape-specific circumstances should we use Indi-seq or Pool-seq? And which populations to select under different landscape scenarios?

Materials and methods

Workflow overview

We first simulated landscapes inhabited by multiple populations of an outbreeding species. For every simulated landscape, migration between adjacent pairs of populations is uniform, and there is one trait of interest which is positively correlated with fitness. This trait is controlled by QTL with purely additive effects. The distribution across the landscape of the corresponding favourable alleles is affected by their population of origin, selection intensity and migration rate. We then simulated a stratified sampling strategy and performed Indi-seq and Pool-seq. Next, GPAS models were built to identify the QTL and predict the phenotypes of individuals and pools. Finally, we assessed how the number of populations sampled, the genotyping strategy (Indi-seq and Pool-seq), selection intensity, migration rate, and the distribution of the favourable alleles across the landscape in relation to the sampled populations affected QTL detection and genomic prediction accuracies.

Landscape simulations

The landscapes were simulated using quantiNemo2 (Neuenschwander et al., 2018), the variables are listed in Table 1, and the fixed parameters are listed in Table S1. We simulated the simple and common mating system consisting of a hermaphrodite species capable of both self- and cross-fertilisation (1/n and 1-1/n probabilities, respectively; n is the population size) with discrete non-overlapping generations. This sufficiently represents allogamous and mixed mating systems, including that of plants with considerable outcrossing, and many animals species.

Variation for a quantitative phenotype over a landscape was simulated as a function of migration rate, number of QTL controlling the trait, causal allele diffusion gradient (the distribution of the favourable alleles across the landscape as it migrates from the populations of origin), and selection intensity with 3 levels for each of these variables (Table 1).

Each landscape consisted of 100 populations arrayed in a uniform square lattice without barriers. Migration was modelled using a 2-dimensional stepping-stone model with bidirectional gene flow with a uniform rate into the 8 adjacent populations and absorbing boundaries.

The quantitative trait was determined by additive QTL with effects sampled from a χ^2 distribution with 1 degree of freedom to generate a cumulative heritability of 0.5. At the initial step of the simulation, all the causal alleles had a frequency (q_0) of 0.01 in the populations of origin and 0 elsewhere. Under the uniform allele diffusion gradient, all populations had q_0=0.01. Under the unidirectional gradient, one boundary row had q_0=0.01 for all 10 populations in that row and 0 elsewhere. Under the bidirectional gradient, two opposite boundary rows had q_0=0.01 in each of the populations and 0 elsewhere. For clarity, these causal allele diffusion gradients are illustrated in Figure 1 Panel A.

Selection was simulated using a generalised logistic model (Richards, 1959) as

$$1/\mathrm{w} = 1 + \mathrm{e}^{y_c - y}, \qquad \text{eq. 1}$$

where *w* is fitness, *y* is the quantitative trait ($y \in \mathbb{R}$), $y_c = y_{min} + s(y_{max}-y_{min})$, with y_{max} and y_{min} the maximum and minimum possible trait values, and *s* is the selection intensity. This selection intensity variable is the minimum relative phenotype value (ranging from 0 to 1) for which

survival rate is at least 50%, assuming that the phenotype is positively correlated with fitness. It follows that a selection intensity of 0.50 or 0.95 means that individuals at the top 50^{th} or 95^{th} percentile, respectively, contribute more to the next generation than those below. We used high selection intensities, i.e. $s \in \{0.5, 0.90, 0.95\}$, to simulate pressing anthropogenic selection pressures, e.g. herbicide or insecticide pressure.

The 10,000 biallelic loci were randomly distributed across a large genome with 7 chromosomes and a total length of 2×10^9 base-pairs and 750 centimorgans. Two hundred generations were simulated to allow the causal alleles to migrate across the landscape and generate the allele frequency distributions specific to each of the three causal allele diffusion gradients. Phenotype values in the final generation were scaled in the 0 to 1 interval and used for the GPAS experiments.

Genome-wide association and trait prediction based on polygenic scores

GPAS was performed on all the populations. Indi-seq and Pool-seq data were assumed to have been generated without genotyping error. We used established tools for Indi-seq data, and developed a suite of tools for Pool-seq data. For Indi-seq data, 384 individuals were sampled per population, simulating four 96-well sample plates or a single 384-well plate commonly used in high-throughput molecular biology workflows. For Pool-seq data, 5 pools per population were sampled, where each pool consisted of 100 individuals. This corresponds to a high power design that was shown to be optimal to capture QTL association (Fournier-Level et al., 2017). GPAS models were trained within each population sampled and cross-validated on all other populations to assess prediction accuracies. For each population, 384 individuals and 5 pools

were used to train and validate the models. This corresponds to an external cross-validation with equally-sized mutually exclusive training and validation sets.

Allele effects were estimated using 6 Indi-seq-based GPAS (Indi-GPAS) and 3 Pool-seq-based GPAS (Pool-GPAS) models. The 6 Indi-GPAS models consisted of: efficient mixed-model association expedited model (EMMAX; Kang et al., 2010), genome-wide complex trait analysis (GCTA; Jiang et al., 2019), and genome-wide efficient mixed-model analysis (GEMMA; Zhou & Stephens, 2012), in combination with 2 types of genetic relationship matrices: GCTA-derived sparse genetic relationship matrix (GRM; off diagonals <0.05 were set to zero; Zaitlen et al., 2013) and GEMMA-derived standardised relatedness matrix (STD; Zhou & Stephens, 2012) (i.e. EMMAX(GRM), EMMAX(STD), GCTA(GRM), GCTA(STD), GEMMA(GRM), and GEMMA(STD)). The 3 Pool-GPAS models consisted of the genome-wide estimation of additive effects based on trait quantile distribution from Pool-seq data (GWAlpha; Fournier-Level et al., 2017), and linear mixed models (LMM) with random pairwise genetic covariance matrix determined by F_{ST} derived using either Hivert's (Hivert et al., 2018) or Weir and Cockerham's method (Weir & Cockerham, 1984) (i.e. GWAlpha, LMM (Fst Hivert), and LMM (Fst Weir & Cockerham)-). The variance components of the LMM were estimated using restricted maximum likelihood.

Phenotype predictions were derived from polygenic scores, i.e. the sum of the products of estimated allele effects and allele dosages for Indi-GPAS or allele frequencies for Pool-GPAS. For the Indi-GPAS models and GWAlpha, this involved a two-step approach. For each training set, the polygenic scores of the training set (s_{train}) were calculated as:

$$s_{\text{train}} = \mathrm{X}_{\text{train}}\beta, \quad \text{eq. 2}$$

where X_{train} is the dosage or frequency of alleles in the training set, and β is the vector of estimated SNP effects. The polygenic scores and actual phenotype values have a linear relationship (Figure S1; mean adjusted R^2=0.97±0.0031 with 1,000 individuals per population for Indi-seq and mean adjusted R^2=0.99±0.0006 with 5 pools per population for Pool-seq) as expected under the additive model used to simulate the phenotypes. These polygenic scores were regressed against the actual phenotype values of the training set $\left(y_{\text{train}}\right)$,

$$y_{\text{train}} = \alpha_0 + \alpha_1 s_{\text{train}}, \quad \text{eq. 3}$$

where α_0 is the intercept, and $\alpha_1$is the slope. The polygenic scores of the validation set, $s_{\text{valid}} = \mathrm{X}_{\text{valid}}\beta$, were transformed into the predicted phenotype values ($y_{\text{predicted}}$) using

$$y_{\text{predicted}} = \alpha_0 + \alpha_1 s_{\text{valid}}. \quad \text{eq. 4}$$

For the Pool-GPAS linear mixed models, the predicted phenotypes ($y_{\text{predicted}}$) were calculated as:

$$y_{\text{predicted}} = \mathrm{X}_{\text{valid}}\beta, \quad \text{eq. 5}$$

where X_{valid} is the matrix of allele frequencies of the validation set, and β is the estimated allelic effects from the GPAS model built using the training set. The trained models were validated on all populations in the landscape.

GPAS performance was measured using three GWAS metrics, and one phenotype prediction metric. The GWAS metrics were:

1 area under the receiver operating curve (AUC) (Fawcett, 2006),

2 true positive rate (TPR) which was defined as the fraction of causal QTL with a significantly associated SNP within 1 kbp, and

3 false positive rate (FPR) which was defined as the fraction of the significantly associated SNPs with no causal QTL within 1 kbp, unless it tags a true QTL through a chain of associated SNPs each less than 1kb apart; multiple associated SNPs within 1kbp were counted as one.

The family-wise type I error rate was set at α=0.05. The metric for phenotype prediction is the root mean square error (RMSE) between actual and predicted phenotype values:

$$\mathrm{RMSE}=\sqrt{\sum \frac{\left(y-\hat{y}\right)^2}{n}}, \qquad \text{eq. 6}$$

where y is the actual phenotypes, $\hat{y}$ is the predicted phenotypes and n is the number of observations.

Sampling strategy optimisation

A total of 405 landscapes were simulated, corresponding to all combinations of the 4 landscape variables with 3 levels each and 5 replicates (Table 1). For each landscape, Indi-GPAS and Pool-GPAS experiments were performed for each population independently. This constitutes the intra-population dataset. The landscape was divided into equally sized rectangular regions, and the approximately central population was selected from each region to simulate a stratified sampling strategy. This is illustrated in Figure 1 Panel B. This constitutes the inter-population dataset. AUC and RMSE were averaged across the populations sampled. TPR and FPR were calculated using the cumulative number of true and false positive candidate loci across the

252 populations sampled. AUC was used to measure the accuracy of QTL detection per population,
253 while TPR and FPR were used to measure QTL detection accuracy of multiple populations.
254
255 The single best performing modelling framework was identified for each genotyping scheme
256 (Indi-GPAS and Pool-GPAS) based on AUC and RMSE for independent populations tests using
257 Tukey's honest significant difference (HSD mean comparison) at $\alpha=0.05$.
258

259 How many populations do we need to sample to yield optimal GPAS performance?

260 To determine how many populations to sample to yield optimal GPAS performance, we used the
261 inter-population dataset. We assessed the suitability of the four metrics (i.e. mean AUC, mean
262 RMSE, TPR and FPR) to address this question by visualising their relationships with the number
263 of populations sampled. Additionally, we compared the expected performance of Indi-GPAS and
264 Pool-GPAS under the same sequencing capacity constraint. The Indi-GPAS experiments we
265 simulated included 384 individuals per population, while Pool-GPAS included only 5 pools per
266 population. Assuming a 5X sequencing depth per individual for Indi-seq (Brouard et al., 2017)
267 and the recommended 50X depth per pool for Pool-seq (Schlötterer et al., 2014), these equate
268 to a sequencing depth of 1,920X per base per population for Indi-seq and only 250X for Pool-
269 seq. This means that for the sequencing capacity required to characterise one population
270 through Indi-seq, approximately 7 populations ($\lfloor 1920/250 \rfloor$) can be characterised through Pool-
271 seq.
272

Under which landscape-specific circumstances should we use Indi-seq or Pool-seq?

The second question we addressed was which landscape-specific circumstances warrant Indi-GPAS or Pool-GPAS? Specifically, if we were to perform GPAS on one population, which sequencing strategy (Indi-seq or Pool-seq) is better, and how does the optimal choice vary with the polygenicity of the trait, selection intensity, and gene flow? The intra-population dataset was analysed using AUC and RMSE as the GPAS performance metrics.

Which populations to select under different landscape scenarios?

To determine which populations to select to best capture the genetic basis of a trait and yield accurate trait predictions under the 3 causal allele diffusion gradients, we analysed the intra-population dataset using AUC and RMSE as the GPAS performance metrics. The populations were classified into 10 groups, where each group represents a row perpendicular to the causal allele diffusion gradient (refer to Figure 1A). The top row corresponds to populations 1 to 10, the second row to populations 11 to 20, and so on. The general and landscape variable-specific trends in GPAS performance across the landscape were visualised using violin plots and means compared using Tukey's HSD (α=0.05). Linear mixed models fitted linear and quadratic relationships (using second degree polynomial fit) between GPAS performance and the row groups. The row group was treated as a numeric variable, and nested within each level of the variables: number of QTL, selection intensity, migration rate, and GPAS model.

Implementation

The landscapes were simulated using quantiNemo2 (Neuenschwander et al., 2018). The genome and QTL information were simulated in R (R Core Team, 2018). The quantiNemo2 outputs were parsed using R and Julia (Nardelli et al., 2018). GEMMA (Zhou & Stephens, 2012), EMMAX (Kang et al., 2010), GCTA (Jiang et al., 2019), and Plink (Purcell et al., 2007) were used for Indi-GPAS. GWAlpha.jl was used for Pool-GPAS. The R package violinplotter was used to generate violin plots with HSD mean comparison grouping. The GNU shell (Free Software Foundation, 2016), Spartan (Lafayette & Wiebelt, 2017), Slurm (Yoo et al., 2003), and GNU parallel (Tange, 2011) were used extensively. The workflow is available in the github repository: https://github.com/jeffersonfparil/GPAS-landscape-simulation.git.

Results

GPAS model selection and the effects of landscape and sampling variables

GEMMA (STD) and GWAlpha showed the best GPAS performances, with >79% AUC and <5.9% RMSE (Table S2). Therefore, these two frameworks were selected as the representatives of Indi-GPAS and Pool-GPAS models, respectively. Overall, Indi-GPAS performed better than Pool-GPAS.

Factors increasing statistical power to identify causal loci through GPAS included a lower number of QTL controlling the trait, more intense selection, higher migration among populations, and more populations sampled (Figure S2). Accuracy in phenotype predictions improved as the number of QTL controlling the trait increases, as selection intensity decreases, and as migration

rate increases (Figure S3). Accuracy is unaffected by the number of populations sampled since each model was trained independently for each population. In addition, power and accuracy are higher when QTL diffuses across the landscape uniformly.

How many populations to sample and when to use Indi-seq or Pool-seq?

TPR and FPR increase logarithmically as the number of populations sampled increases (Figure 2), so there is no optimum based on these metrics.

Indi-GPAS achieves greater power than Pool-GPAS at the cost of a higher false positive rate (Figure 2). However, Pool-GPAS can outperform Indi-GPAS under the assumptions detailed in the materials and methods section, where for every population characterised with Indi-seq, approximately 7 populations can be characterised with Pool-seq. Under this 1:7 ratio, Indi-GPAS on 10 populations yield an average TPR of 0.388 and FPR of 0.0150; for the same sequencing capacity Pool-GPAS can be performed on 70 populations, yielding an average TPR of 0.418 and FPR of 0.0115. We explored a range of ratios deviating from this 1:7 ratio. This is because the 5X depth requirement for variant calling in Indi-seq and 50X depth for allele frequency estimation in Pool-seq depend on the species of interest and the resources available. Lower ratios, e.g. 1:8 to 1:10, mean even more populations can be characterised with Pool-seq for every population characterised with Indi-seq. Using our simulated data to explore various ratios, we find that there exists a range where Pool-GPAS can outperform Indi-GPAS, i.e. TPR is higher and FPR is lower for Pool-GPAS than Indi-GPAS (Figure 3). This shows that characterising more of the landscape at low resolution can be better than characterising a small portion of the landscape at high resolution.

If we were to perform GPAS on one population, Indi-GPAS is better than Pool-GPAS. However in cases where selection intensity is high (i.e. 0.90 to 0.95) or migration rate is high (i.e. 0.01) Pool-GPAS performance is not significantly different from Indi-GPAS in terms of prediction accuracy (Figure 4).

Which populations to select under different landscape scenarios?

GPAS performance is maximised in populations with high genetic variability which at the landscape level, means sampled close to the place of origin of the causal allele (Figure 5). This area of high genetic variability is characterised by intermediate causal allele frequencies which translate into populations with high phenotypic variability. In the absence of a causal allele diffusion gradient (i.e. uniform causal allele distribution), no row seems to be optimal for sampling, except for some slightly better performance from populations in the middle rows. Under unidirectional gradient (i.e. causal alleles originated from the top row and diffused downwards hence a single diffusion front) and in terms of QTL detection accuracy, sampling the populations from the top row is optimal; however, in terms of prediction accuracy, the populations in the middle rows appear to be better. Under bidirectional gradient (i.e. causal alleles originated from the top and bottom rows hence two diffusion fronts) both QTL detection and prediction accuracies are optimal in the populations from the top and bottom rows. These trends across the landscape correlate with the trends in the mean number of polymorphic QTL per population and causal allele frequencies.

In the presence of causal allele diffusion gradients, the relationship of the sampling location (defined as rows perpendicular to the diffusion gradient) with both QTL detection and prediction accuracies appears to be quadratic, except for QTL detection accuracy under unidirectional causal allele diffusion, for which the relationship is linear (Figure 5). In terms of GWAS accuracy as measured by AUC, sampling near the diffusion fronts becomes less important (i.e. slope under unidirectional gradient and curvature under bidirectional gradient are reduced) as the number of QTL increases, as selection intensity decreases, and as migration rate increases (Figure 6 columns 1-3). In addition, sampling near the diffusion fronts is more important for Pool-GPAS than Indi-GPAS (Figure 6 column 4), in other words, power diminishes quicker for Pool-GPAS than Indi-GPAS as we move away from areas of high diversity.

In terms of prediction accuracy as measured by RMSE, the degree to which the middle rows (i.e. areas of high genetic and phenotypic variability) are the optimal sampling locations under unidirectional diffusion decreases (i.e. curvature becomes less severe) as the number of QTL increases, as selection intensity decreases, and as migration rate increases (Figure 7 top graphs). Also, sampling from the middle rows under unidirectional diffusion is slightly more important for Indi-GPAS than Pool-GPAS. On the other hand, the degree to which the top and bottom rows are optimal under bidirectional diffusion decreases (i.e. curvature becomes less severe) as the number of QTL, selection intensity, and migration rate increase (Figure 7 bottom graphs). Also, sampling from the top and bottom rows under bidirectional diffusion is more important for Pool-GPAS than Indi-GPAS, in other words, similar to that of power, accuracy diminishes quicker for Pool-GPAS than Indi-GPAS as we move away from areas of high diversity.

The trends in GPAS performance across the landscape correlate with the trends in genetic variability (expressed in terms of causal allele frequency, i.e. frequencies closer to 0.5 indicates higher diversity; Figures S4 to S9). Opposite trends are observed between causal allele diffusion gradients for RMSE as selection intensity increases, i.e. in the middle of the landscape, RMSE is minimised for unidirectional causal diffusion but maximised for bidirectional causal allele diffusion (Figure S5 and S8).

Discussion

GPAS has the potential to extend the scope of genomic studies in ecology and evolution beyond environment association and niche modelling (Dormann et al., 2012; Exposito-Alonso et al., 2018; Fournier-Level et al., 2011). The predicted environmental range of individual genotypes determined through genome-environment associations (Manel et al., 2018) can be complemented by the phenotype predictions of GPAS (Cotto et al., 2017). This can be transformational for the way we monitor invasive species or assess the adaptive potential of endangered ones. We provide recommendations for optimising the sampling strategy to maximise the power to detect QTL and the accuracy of quantitative phenotype prediction in natural populations of any outbreeding species. We stress the importance of capturing sufficient representation of the genetic variability present over the landscape by sampling populations from areas of high genetic diversity. On a per population basis or if only one population were to be sampled, we recommend using Indi-seq over Pool-seq. We have not considered phenotyping costs here, but if it is high, then it would increase the attractiveness of Indi-seq to maximise

information per unit cost. However, similar to a study on the estimation of population differentiation (Goudet & Büchi, 2006) and a meta analysis of several landscape genomics studies (Santos & Gaiotto, 2020), w demonstrated the value of shallow but extensive genotyping with Pool-seq in maximising the number of populations that can be analysed without compromising power. This is especially true if the aim is to predict phenotypes of some future populations for the rapid and timely monitoring of invasive and threatened species.

How many populations to sample?

Our results emphasise the need to sample as many populations as possible from regions of high genetic diversity. The power to detect QTL is maximised if all the populations in the landscape were included in the study. This is possible for endangered species with a small number of populations in the wild. However this is not feasible for species with a healthier number of populations. The best populations to sample are located in areas of high genetic diversity which manifests as areas with high trait variability where the causal alleles are at intermediate frequencies.

Our results show a diminishing return in terms of GPAS power when increasing the number of populations sampled. This is consistent with a study by Selmoni et al. (2020) on sampling strategy optimisation for landscape genomics which found that sampling an intermediate number of sites can perform as well as maximising the number of sites sampled. This study only considered Indi-seq and tested different sample sizes per population, and our approach is comparable because using Indi-seq equates to a high-resolution characterisation of the landscape and Pool-seq to a low-resolution one. Our analysis extends this result further because

it is independent of the number of individuals sampled per population. The cost-effectiveness of Pool-seq allows for more populations to be sampled and included in the study than Indi-seq.

When to use Indi-seq or Pool-seq?

The number of individuals per population selected for our Indi-GPAS simulations (384 individuals) exceeds the sample size of most ecological studies (e.g. 100-200 individual samples per population in birds (Hansson et al., 2018; Perrier et al., 2018), <100 samples per population in trees (Cappa et al., 2013; Holliday et al., 2010), ~100 samples per population in mammals (Johnston et al., 2011; Pallares et al., 2014), and <20 samples per population in fish (Willing et al., 2010)). Thus for most experiments, the power of Indi-GPAS is expected to be lower than in our simulations. On the contrary, the power of Pool-GPAS is expected to remain the same since five pools per population was found to be optimal (Fournier-Level et al., 2017) and each pool can include a non-limiting number of individuals. The sequencing capacity required for Indi-seq is always higher than for Pool-seq, and more populations can be characterised with Pool-seq than with Indi-seq under the same budgetary constraints (Schlötterer et al., 2014). Therefore, the range of the number of populations sampled where Pool-GPAS outperforms Indi-GPAS is likely to be even broader than reported here, as long as the genomic characterisation approach yields accurate genomic data.

Sufficient depth of coverage is required to correct for sequencing errors inherent to current high-throughput sequencing technologies. Indi-seq can provide high-resolution genomic information for a population, but comes at a high cost. A given genomic region needs to be sequenced at least 5 times for each individual to correct for sequencing errors and yield accurate basecalling

information (Brouard et al., 2017). Additionally, many individuals are required to accurately represent a population. This is only resource-effective when the individuals are part of an association panel and the genomic information can be leveraged across several research projects (Robin et al., 2019). On the other hand, Pool-seq generates low-resolution genomic information on a population that is cost-effective while maintaining high power. To maximise the accuracy of allele frequency estimation (i.e. to correct for sequencing errors and sufficiently represent the pool), Pool-seq guidelines and recommendations have been proposed (Schlötterer et al. 2014; Fracassetti et al. 2015; Anand et al. 2016). Pool sizes of at least 25 (Gautier et al. 2013; Fracassetti et al. 2015) to 40 individuals (Schlötterer et al. 2014), and depths greater than 50X (Zhu et al. 2012) to 65X (Gautier et al. 2013) have been recommended. Hundreds of individuals can be pooled to yield accurate allele frequency data (Schlötterer et al., 2014). This means that in an outbreeding species, a few pools consisting of hundreds of individuals each can represent a population better than a few individuals.

Pool-seq is more widely used than Indi-seq in ecological and evolutionary studies because the focus is generally on populations rather than individuals, and because of its cost-effectiveness (Futschik & Schlötterer, 2010). In contrast, Cutler and Jensen (2010) concluded that Indi-seq should be preferred over Pool-seq for many applications due to the loss of haplotype and LD information. They focused on applications in human and model organisms, whereas Pool-seq has its highest impact for high-throughput data acquisition in non-model species of critical ecological and economical importance.

Which populations to select under different landscape scenarios?

We have shown that sampling from genetically diverse populations maximises GPAS performance. Capturing greater genetic diversity was shown to increase the power to detect causal loci (Alqudah et al., 2020; Rosenberg et al., 2010; Wojcik et al., 2019). Similarly, populations which represent the overall diversity found in the landscape or are similar to the validation populations, improve prediction accuracies (Akdemir & Isidro-Sánchez, 2019; Asoro et al., 2011; Edwards et al., 2019). Populations with high genetic diversity were found along the diffusion fronts, i.e. the areas where the causal alleles migrate from their site of origin into the neighbouring populations. The rate at which GPAS performance decreases as we sample farther away from the diffusion fronts correlates with the decrease in genetic diversity at the causal loci. In the absence of prior genomic information, the areas of high genetic diversity coincide with regions of high phenotypic diversity. Gaining prior information on the location of these areas of high genetic diversity and causal allele diffusion fronts or more broadly the landscape of adaptive genetic diversity (Eckert & Dyer, 2012) is key to an optimal sampling strategy.

When the causal allele diffusion gradient is unknown and a uniform causal allele distribution is assumed, there is a small advantage in choosing populations in the middle of the landscape. This can be explained by the absorbing boundaries used in the migration model which simulates a restricted range whereby alleles going beyond the border are lost. This reflects the phenomenon in fringe populations where migration regularly occurs beyond the suitable environmental niche of the species and the migrants fail to survive (Sexton et al., 2009). This is expected to apply to organisms with restricted range such as corals (Guan et al., 2015), 24% of

coral reef fish species (Hawkins et al., 2006), and organisms in hydrothermal vents (Mullineaux et al., 2018). However, in cases where non-uniform causal allele distribution is assumed (e.g. temperature response in lodgepole pine and interior spruce (Liepe et al., 2015); and white spruce (Hornoy et al., 2015)), these populations along the borders of the species range are important to sample because they may carry unique advantageous variants, especially in the context of climate change adaptation (Geber & Eckhart, 2005; Bridle & Vines, 2007; le Roux , 2009; Pais et al., 2020).

When considering the genetic architecture of a trait, there is less power to detect QTL but higher prediction accuracy for highly polygenic traits than for traits controlled by fewer loci. As a consequence, if we expect the polygenic trait to be controlled by relatively few loci, and we are more focused on mapping its genetic basis than on phenotype prediction, then sufficient power to detect QTL can be achieved with less populations sampled. Biotic stress resistance traits were often shown to be oligogenic, for example resistance to *Pseudomonas syringae* in *Arabidopsis thaliana* (Atwell et al., 2010), and resistance to *Cronartium ribicola* in sugar pine (Weiss et al., 2020). On the other hand, if we expect the trait to be highly polygenic, and we are more focused on phenotype prediction for monitoring purposes, then we have to sample as many populations as possible to sufficiently represent the variation across the landscape and maximise prediction accuracy. This is typically the case for abiotic stress resistance traits such as aluminium and proton tolerance in *Arabidopsis thaliana* (Nakano, et al., 2020), and coral bleaching resistance in *Acropora millepora* (Fuller et al., 2020) As the number of loci controlling the trait increases, the selection pressure acting on each locus decreases (Walsh & Lynch, 2018). This reduces the power to detect QTL since the individual contribution of each QTL

519 decreases as more loci control the trait (Wang & Xu, 2019). This in turn reduces the proportion
520 of polymorphic QTL within populations: if the majority of the QTL have small effects, they have a
521 higher chance of getting lost due to drift than QTL with large effects. On the other hand,
522 prediction accuracy increases since the rate at which genetic variance decreases due to
523 directional selection is reduced as the number of QTL increases. Genetic variance should
524 eventually become zero under constant stabilising or directional selection, but the rate of this
525 reduction becomes slower with an increased number of loci controlling the trait (Crow & Kimura,
526 1970). Hence, GWAS and GP complement each other to achieve high QTL detection accuracy
527 or high prediction accuracy for quantitative traits controlled by any number of loci.
528
529 When considering selection intensity, there is higher power to detect QTL but lower prediction
530 accuracy in populations under intense selection. As a consequence, populations under high
531 selection pressure can be selected for high-power GWAS: for example, weed and insect pest
532 populations in agricultural areas including herbicide-resistant *Lolium rigidum* (Powles et al.,
533 1998), and Bt-resistant *Helicoverpa armigera* (Jin et al., 2018). On the other hand, populations
534 under low selection pressure can be selected for GP. For example, non-target organisms
535 including *Drosophila melanogaster* populations which are resistant to Bt toxins (Babin et al.,
536 2020) and imidacloprid (Fournier-Level et al., 2019) Increasing the selection intensity increases
537 QTL detection power, since the effect of individual QTL becomes greater within each population
538 (Wang & Xu, 2019) and less QTL alleles are lost due to drift. However, the predictive ability will
539 be reduced by the Bulmer effect (Bulmer, 1971), where covariance between loci (partially
540 explained by linkage disequilibrium (Walsh & Lynch, 2018)) is reduced after selection. Increasing
541 selection intensity magnifies this reduction resulting in diminished additive genetic variance and

less predictive models. This further solidifies the complementary nature of GWAS and GP and the utility of performing both with GPAS.

When considering migration rate, there is more power to detect QTL and greater prediction accuracy in populations experiencing high migration than in reproductively isolated ones. As a consequence, if gene flow is high, then less populations need to be sampled to sufficiently represent the variation across the landscape. This is the case for highly mobile organisms including birds (Pulido, 2007); and fishes (Brodersen et al., 2008), On the other hand, if gene flow is low which results in highly structured landscapes, then more populations need to be sampled to sufficiently represent the landscape variation. This can be the case in landscapes with considerable natural or artificial barriers to migration, and in species with reproductive structures impeding outbreeding, e.g. cleistogamous plants including *Crotalaria micans* (Etcheverry et al., 2003) and *Vigna caracalla* (Etcheverry et al., 2008). Increasing migration rate decreases differentiation between populations allowing for more causal alleles to be shared This leads to higher additive genetic variance per population, resulting in higher power and more accurate predictions (Liu et al., 2020).

Conclusion

Genome-wide association studies and genomic prediction (GPAS) are poised to complement existing methodologies in ecology in evolution. GPAS provides powerful tools to dissect the genetic basis of ecologically important quantitative traits including fitness and to rapidly monitor natural populations including invasive and threatened species. Understanding how the number

564 of population samples and the different landscape properties affect the QTL detection power and
565 phenotype prediction accuracies is integral to planning population collections for GPAS
566 experiments. We recommend sampling as many populations as possible from areas of high
567 genetic diversity. We also recommend Pool-seq whenever Indi-seq is too costly; since sampling
568 more populations at the cost of lower resolution can be better than characterising a small
569 number of populations at high resolution. The complementary nature of GWAS and GP allows
570 good QTL detection power or prediction accuracy under low to high trait polygenicity and
571 selection intensity. In the absence of prior information on the areas of high genetic diversity, we
572 recommend against sampling populations at the border of the species' range.
573

Acknowledgements

The authors wish to thank Uli Felzmann from IT services, Faculty of Science, The University of Melbourne for his outstanding support. We acknowledge the extensive use of the Spartan High Performance Computing and the Melbourne Research Cloud systems. Part of this work was supported through the Computational Biology Research Initiative Seed Fund awarded to AF-L, and the David Hay Fund.

Data accessibility

Codes to reproduce the simulation data and the subsequent analysis are publicly available on github: https://github.com/jeffersonfparil/GPAS-landscape-simulation.

Author Contributions

JFP, DJB and AFL designed the study. JFP analysed the data. JFP and AFL wrote the manuscript with input from DJB.

References

Akdemir, D., & Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports*, *9*(1), 1446. doi: 10.1038/s41598-018-38081-6

Alqudah, A. M., Sallam, A., Stephen Baenziger, P., & Börner, A. (2020). GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley – A review. *Journal of Advanced Research*, *22*, 119–135. doi: 10.1016/j.jare.2019.10.013

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., & Jannink, J.-L. (2011). Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. *The Plant Genome*, *4*(2), 132–144. doi: 10.3835/plantgenome2011.02.0007

Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. (2010). "Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines." *Nature* 465, no. 7298: 627–31. https://doi.org/10.1038/nature08800.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, *3*(10), e3376. doi: 10.1371/journal.pone.0003376

Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stöbe, P., Futschik, A., & Schlötterer, C. (2013). A Genome-Wide, Fine-Scale Map of Natural Pigmentation Variation in Drosophila melanogaster. *PLoS Genetics*, *9*(6), e1003534. doi: 10.1371/journal.pgen.1003534

Beissinger, T. M., Hirsch, C. N., Vaillancourt, B., Deshpande, S., Barry, K., Buell, C. R., … de Leon, N. (2014). A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. *Genetics*, *196*(3), 829–40. doi: 10.1534/genetics.113.160655

Bélanger, S., Esteves, P., Clermont, I., Jean, M., & Belzile, F. (2016). Genotyping-by-Sequencing on Pooled Samples and its Use in Measuring Segregation Bias during the Course of Androgenesis in Barley. *The Plant Genome*, *9*(1), 0. doi: 10.3835/plantgenome2014.10.0073

Bernal-Vasquez, AM., Möhring, J., Schmidt, M., Schönleben, M., Schön, S.-C., & Piepho, H.-P. (2014). The importance of phenotypic data analysis for genomic prediction - a case study comparing different spatial models in rye. *BMC Genomics 15, 646* . https://doi.org/10.1186/1471-2164-15-646

Boitard, S., Schlotterer, C., Nolte, V., Pandey, R. V., & Futschik, A. (2012). Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. *Molecular Biology and*

633 *Evolution*, *29*(9), 2177–2186. doi: 10.1093/molbev/mss090
634
635 Bridle, Jon R., & Vines, Timothy H., (2007). Limits to evolution at range margins: when and why
636 does adaptation fail?. *Trends in Ecology & Evolution. Volume 22, Issue 3*, pp 140-147.
637 ISSN 0169-5347. https://doi.org/10.1016/j.tree.2006.11.002.
638
639 Brodersen, Jakob, P. Anders Nilsson, Lars-Anders Hansson, Christian Skov, & Christer
640 Brönmark. (2008). "Condition-Dependent Individual Decision-Making Determines
641 Cyprinid Partial Migration." *Ecology* 89, no. 5: 1195–1200. https://doi.org/10.1890/07-
642 1318.1.
643
644 Brouard, J.-S., Boyle, B., Ibeagha-Awemu, E. M., & Bissonnette, N. (2017). Low-depth
645 genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the
646 selection of high quality genotypes and the accuracy of imputation. *BMC Genetics*, *18*.
647 doi: 10.1186/s12863-017-0501-y
648
649 Bulmer, M. G. (1971). The Effect of Selection on Genetic Variability. *The American Naturalist*.
650 Retrieved from https://www.journals.uchicago.edu/doi/10.1086/282718
651
652 Bustos-Korts, D., Malosetti, M., Chapman S., Biddulph, B., & van Eeuwijk, F. (2016).
653 Improvement of Predictive Ability by Uniform Coverage of the Target Genetic Space. *G3*
654 *Genes*|*Genomes*|*Genetics, Volume 6, Issue 11,* 1 November 2016, Pages 3733–3747,
655 https://doi.org/10.1534/g3.116.035410
656 Cappa, E. P., El-Kassaby, Y. A., Garcia, M. N., Acuña, C., Borralho, N. M. G., Grattapaglia, D.,
657 & Poltri, S. N. M. (2013). Impacts of Population Structure and Analytical Models in
658 Genome-Wide Association Studies of Complex Traits in Forest Trees: A Case Study in
659 Eucalyptus globulus. *PLOS ONE*, *8*(11), e81267. doi: 10.1371/journal.pone.0081267
660
661 Cheng, C., White, B. J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M. W., & Besansky, N.
662 J. (2012). Ecological genomics of anopheles gambiae along a latitudinal cline: A
663 population-resequencing approach. *Genetics*, *190*(4), 1417–1432. doi:
664 10.1534/genetics.111.137794
665
666 Cotto, O., Wessely, J., Georges, D., Klonner, G., Schmid, M., Dullinger, S., Thuiller, W., &
667 Guillaume, F.. (2017). A dynamic eco-evolutionary model predicts slow response of
668 alpine plants to climate warming. *Nat Commun* **8,** 15399.
669 https://doi.org/10.1038/ncomms15399
670
671 Crow, J. F., & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. doi:
672 10.2307/1529706
673
674 Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, *186*(1), 41–3. doi:
675 10.1534/genetics.110.121012
676
677 de Roos, A. P. W., Hayes, B. J., & Goddard, M. E. (2009). Reliability of genomic predictions
678 across multiple populations. *Genetics*, *183*(4), 1545–53. doi:

679 10.1534/genetics.109.104935
680
681 Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., … Singer, A.
682 (2012). Correlation and process in species distribution models: bridging a dichotomy.
683 *Journal of Biogeography*, *39*(12), 2119–2131. doi: 10.1111/j.1365-2699.2011.02659.x
684
685 Eckert, A. J., & Dyer, R. J. (2012). Defining the landscape of adaptive genetic diversity.
686 Molecular Ecology, 21(12), 2836–2838. doi: 10.1111/j.1365-294X.2012.05615.x
687
688 Edwards, S. M., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., … Hickey, J. M.
689 (2019). The effects of training population design on genomic prediction accuracy in
690 wheat. *Theoretical and Applied Genetics*, *132*(7), 1943–1952. doi: 10.1007/s00122-019-
691 03327-y
692
693 Etcheverry, Angela Virginia, Maria Mercedes Alemán, and Trinidad Figueroa Fleming. (2008).
694 “Flower Morphology, Pollination Biology and Mating System of the Complex Flower of
695 *Vigna caracalla* (Fabaceae: Papilionoideae).” *Annals of Botany* 102, no. 3 (September 1,
696 2008): 305–16. https://doi.org/10.1093/aob/mcn106.
697
698 Etcheverry, A. V., J. J. Protomastro, and C. Westerkamp. (2003). “Delayed Autonomous Self-
699 Pollination in the Colonizer Crotalaria Micans (Fabaceae: Papilionoideae): Structural and
700 Functional Aspects.” *Plant Systematics and Evolution* 239, no. 1/2 (2003): 15–28.
701
702 Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H. A., & Weigel, D. (2018).
703 Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis
704 thaliana. *Nature Ecology & Evolution*, *2*(2), 352–358. doi: 10.1038/s41559-017-0423-0
705
706 Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C., … SanCristobal, M.
707 (2017). Accounting for linkage disequilibrium in genome scans for selection without
708 individual genotypes: The local score approach. *Molecular Ecology*, *26*(14), 3700–3714.
709 doi: 10.1111/mec.14141
710
711 Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–
712 874. doi: 10.1016/j.patrec.2005.10.010
713
714 Fournier-Level, A., Good, R.T., Wilcox, S.A. Rane, R.V., Schiffer, M., Chen, W., Battlay, P.,
715 Perry, T., Batterham, P., Hoffmann, A.A., & Robin, C. (2019). The spread of resistance to
716 imidacloprid is restricted by thermotolerance in natural populations of *Drosophila*
717 *melanogaster*. Nat Ecol Evol 3, 647–656. https://doi.org/10.1038/s41559-019-0837-y
718
719 Fournier-Level, A., Korte, A., Cooper, M. D., Nordborg, M., Schmitt, J., & Wilczek, A. M. (2011).
720 A Map of Local Adaptation in Arabidopsis thaliana. *Science*, *334*(6052), 86–89. doi:
721 10.1126/science.1209271
722
723 Fournier-Level, Alexandre, Robin, C., & Balding, D. J. (2017). GWAlpha: Genome-wide
724 estimation of additive effects (alpha) based on trait quantile distribution from pool-

sequencing experiments. *Bioinformatics*. doi: 10.1093/bioinformatics/btw805

Fracassetti, M., Griffin, P. C., & Willi, Y. (2015). Validation of Pooled Whole-Genome Re-Sequencing in Arabidopsis lyrata. *PloS One*, *10*(10), e0140462. doi: 10.1371/journal.pone.0140462

Free Software Foundation. (2016). *GNU bash*. Retrieved from https://www.gnu.org/software/bash/

Fuller, Zachary L., Veronique J. L. Mocellin, Luke A. Morris, Neal Cantin, Jihanne Shepherd, Luke Sarre, Julie Peng, et al. (2020) “Population Genetics of the Coral Acropora Millepora: Toward Genomic Prediction of Bleaching.” *Science* 369, no. 6501 (July 17, 2020). https://doi.org/10.1126/science.aba4674.

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, *186*(1), 207–18. doi: 10.1534/genetics.110.114397

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., … Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), 3766–3779. doi: 10.1111/mec.12360

Geber, M.A. & Eckhart, V.M. (2005). Experimental studies of adaptation in *Clarkia xantiana*: II. Fitness variation across a subspecies border. *Evolution, 59: 521-531.* https://doi.org/10.1111/j.0014-3820.2005.tb01012.x

Gondro, C., van der Werf, J., & Hayes, B. (Eds.). (2013). *Genome-wide association studies and genomic prediction* (1st ed.). Humana Press. Doi: 10.1007/978-1-62703-447-0_20

Goudet, J., & Büchi. (2006). The effects of dominance, regular inbreeding and sampling design on QST, an estimator of population differentiation for quantitative traits. *Genetics 172*, 1337-1347. https://doi.org/10.1534/genetics.105.050583.

Guan, Y., Hohn, S., & Merico, A. (2015). Suitable environmental ranges for potential coral reef habitats in the tropical ocean. *PloS one*, *10*(6), e0128831. https://doi.org/10.1371/journal.pone.0128831

Guo, X., Cericola, F., Fè, D., Pedersen, M. G., Lenk, I., Jensen, C. S., … Janss, L. L. (2018). Genomic Prediction in Tetraploid Ryegrass Using Allele Frequencies Based on Genotyping by Sequencing. *Frontiers in Plant Science*, *9*, 1165. doi: 10.3389/fpls.2018.01165

Gurdasani, D., Barroso, I., Zeggini, E. et al. (2019). Genomics of disease risk in globally diverse populations. *Nat Rev Genet 20,* 520–535. https://doi.org/10.1038/s41576-019-0144-0

Hansson, B., Sigeman, H., Stervander, M., Tarka, M., Ponnikas, S., Strandh, M., … Hasselquist, D. (2018). Contrasting results from GWAS and QTL mapping on wing length in great reed warblers. *Molecular Ecology Resources*, *18*(4), 867–876. doi: 10.1111/1755-0998.12785

Hawkins, J.P., Roberts, C.M. & Clark, V. (2000), The threatened status of restricted−range coral reef fish species. *Animal Conservation*, 3: 81-88. https://doi.org/10.1111/j.1469-1795.2000.tb00089.x

Hornoy, Benjamin, Nathalie Pavy, Sébastien Gérardi, Jean Beaulieu, & Jean Bousquet. (2015). "Genetic Adaptation to Climate in White Spruce Involves Small to Moderate Allele Frequency Shifts in Functionally Diverse Genes." *Genome Biology and Evolution* 7, no. 12: 3269–85. https://doi.org/10.1093/gbe/evv218.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring Genetic Differentiation from Pool-seq Data. *Genetics*, *210*(1), 315–330. doi: 10.1534/genetics.118.300900

Hoel, P. (1943). The Accuracy of Sampling Methods in Ecology on JSTOR. *The Annals of Mathematical Statistics*, *14*(3), 289–300.

Holliday, J. A., Ritland, K., & Aitken, S. N. (2010). Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (Picea sitchensis). *New Phytologist*, *188*(2), 501–514. doi: 10.1111/j.1469-8137.2010.03380.x

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12), 1749–1755. doi: 10.1038/s41588-019-0530-8

Johnston, S. E., McEWAN, J. C., Pickering, N. K., Kijas, J. W., Beraldi, D., Pilkington, J. G., … Slate, J. (2011). Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Molecular Ecology*, *20*(12), 2555–2566. doi: 10.1111/j.1365-294X.2011.05076.x

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., … Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. doi: 10.1038/ng.548

Knight, J., Saccone, S. F., Zhang, Z., Ballinger, D. G., & Rice, J. P. (2009). A Comparison of Association Statistics between Pooled and Individual Genotypes. *Human Heredity*, *67*(4), 219–225. doi: 10.1159/000194975

Lafayette, L., & Wiebelt, B. (2017). Spartan and NEMO: Two HPC-Cloud Hybrid Implementations. *2017 IEEE 13th International Conference on E-Science (e-Science)*,

458–459. Auckland: IEEE. doi: 10.1109/eScience.2017.70

Le Roux P. (2009) Plants at the margin. Ecological limits and climate change. *Ann Bot. 104(7):ix*. doi:10.1093/aob/mcp220

Li, Y., Zhang, X.-X., Mao, R.-L., Yang, J., Miao, C.-Y., Li, Z., & Qiu, Y.-X. (2017). Ten Years of Landscape Genomics: Challenges and Opportunities. *Frontiers in Plant Science*, *8*, 2136. doi: 10.3389/fpls.2017.02136

Liepe, Katharina J., Andreas Hamann, Pia Smets, Connor R. Fitzpatrick, & Sally N. Aitken. (2016). “Adaptation of Lodgepole Pine and Interior Spruce to Climate: Implications for Reforestation in a Warming World.” *Evolutionary Applications* 9, no. 2: 409–19. https://doi.org/10.1111/eva.12345.

Liu, L., Wang, Y., Zhang, D., Chen, Z., Chen, X., Su, Z., & He, X. (2020). The Origin of Additive Genetic Variance Driven by Positive Selection. *Molecular Biology and Evolution*, *37*(8), 2300–2308. doi: 10.1093/molbev/msaa085

Long, A. D., & Langley, C. H. (1999). The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits. *Genome Research*, *9*(8), 720–731. doi: 10.1101/gr.9.8.720

Macgregor, S., Visscher, P. M., & Montgomery, G. (2006). Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates. *Nucleic Acids Research*, *34*(7). doi: 10.1093/nar/gkl136

Manel, S., Andrello, M., Henry, K., Verdelet, D., Darracq, A., Guerin, P.-E., … Devaux, P. (2018). Predicting genotype environmental range from genome–environment associations. Molecular Ecology, 27(13), 2823–2833. doi: 10.1111/mec.14723

Micheletti, S. J., & Narum, S. R. (2018). Utility of pooled sequencing for association mapping in nonmodel organisms. *Molecular Ecology Resources*, *18*(4), 825–837. doi: 10.1111/1755-0998.12784

Mullineaux, Lauren S., Anna Metaxas, Stace E. Beaulieu, Monika Bright, Sabine Gollner, Benjamin M. Grupe, Santiago Herrera, et al. (2018). “Exploring the Ecology of Deep-Sea Hydrothermal Vents in a Metacommunity Framework.” *Frontiers in Marine Science* 5. https://doi.org/10.3389/fmars.2018.00049.

Neuenschwander, S., Michaud, F., Goudet, J., & Stegle, O. (2018). quantiNemo 2: a swiss knife to simulate complex demographic and genetic scenarios, forward and backward in time. *Bioinformatics*. doi: 10.1093/bioinformatics/bty737

Nielsen, E. S., Henriques, R., Toonen, R. J., Knapp, I. S. S., Guo, B., & von der Heyden, S. (2018). Complex signatures of genomic variation of two non-model marine species in a homogeneous environment. *BMC Genomics*, *19*(1), 347. doi: 10.1186/s12864-018-4721-

y

Pais, A.L., Whetten, R.W. & Xiang, Q.−Y. (2020). Population structure, landscape genomics, and genetic signatures of adaptation to exotic disease pressure in Cornus florida L. - Insights from GWAS and GBS data. *J. Syst. Evol., 58*: 546-570. https://doi.org/10.1111/jse.12592

Pallares, L. F., Harr, B., Turner, L. M., & Tautz, D. (2014). Use of a natural hybrid zone for genomewide association mapping of craniofacial traits in the house mouse. *Molecular Ecology*, *23*(23), 5756–5770. doi: 10.1111/mec.12968

Perrier, C., Delahaie, B., & Charmantier, A. (2018). Heritability estimates from genomewide relatedness matrices in wild populations: Application to a passerine, using a small sample size. *Molecular Ecology Resources*, *18*(4), 838–853. doi: 10.1111/1755-0998.12886

Powles, Stephen B, Debrah F Lorraine-Colwill, James J Dellow, & Christopher Preston. (1998). "Evolved Resistance to Glyphosate in Rigid Ryegrass (*Lolium rigidum*) in Australia." *Weed Science* 46, no. 5: 604–7.

Pulido, Francisco. (2007) "The Genetics and Evolution of Avian Migration." *BioScience* 57, no. 2: 165–74. https://doi.org/10.1641/B570211.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. doi: 10.1086/519795

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from https://www.r-project.org/

Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., & Fischer, M. C. (2013). Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species. *PLoS ONE*, *8*(11), e80422. doi: 10.1371/journal.pone.0080422

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*, *10*(2), 290–301. doi: 10.1093/jxb/10.2.290

Riginos, C., Crandall, E.D., Liggins, L., Bongaerts, P., & Treml, E.A. (2016). Navigating the currents of seascape genomics: how spatial analyses can augment population genomic studies. *Current Zoology, Volume 62, Issue 6*, December 2016, Pages 581–601, https://doi.org/10.1093/cz/zow067

Rincent, R., Charcosset, A., & Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured

populations. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, *130*(11), 2231–2247. https://doi.org/10.1007/s00122-017-2956-7

Robin, C., Battlay, P., & Fournier-Level, A. (2019). What can genetic association panels tell us about evolutionary processes in insects? *Current Opinion in Insect Science*, *31*, 99–105. doi: 10.1016/j.cois.2018.12.004

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., & Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, *11*(5), 356–366. doi: 10.1038/nrg2760

Santos, A. S., & Gaiotto, F. A. (2020). Knowledge status and sampling strategies to maximize cost-benefit ratio of studies in landscape genomics of wild plants. *Scientific Reports*, *10*(1), 1–9. doi: 10.1038/s41598-020-60788-8

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature Reviews*, *15*, 749–765.

Selmoni, O., Vajana, E., Guillaume, A., Rochat, E., & Joost, S. (2020). Sampling strategy optimization to increase statistical power in landscape genomics: A simulation-based approach. *Molecular Ecology Resources*, *20*(1), 154–169. doi: 10.1111/1755-0998.13095

Sexton, J. P., McIntyre, P. J., Angert, A. L., & Rice, K. J. (2009). Evolution and Ecology of Species Range Limits. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 415–436. doi: 10.1146/annurev.ecolsys.110308.120317

Tange, O. (2011). *GNU Parallel - The Command-Line Power Tool*. The USENIX Magazine.

Visscher, P.M., Brown, M.A., McCarthy, M.I., & Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet. 90.* pp 7-24

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics. Volume 101, Issue 1.* pp 5-22, ISSN 0002-9297. https://doi.org/10.1016/j.ajhg.2017.06.005.

Walsh, B., & Lynch, M. (2018). *Evolution and selection of quantitative traits* (1st ed.). Oxford University Press. Retrieved from http://nitro.biosci.arizona.edu/zbook/NewVolume_2/newvol2.html#2B

Wang, M., & Xu, S. (2019). Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity*, *123*(3), 287–306. doi: 10.1038/s41437-019-0205-3

Ward, Sarah M., & Marie Jasieniuk. (2009). Review: Sampling Weedy and Invasive Plant

Populations for Genetic Diversity Analysis. *Weed Science 57, no. 6* : 593-602. http://www.jstor.org/stable/40586878.

Weiss, Matthew, Richard A. Sniezko, Daniela Puiu, Marc W. Crepeau, Kristian Stevens, Steven L. Salzberg, Charles H. Langley, David B. Neale, and Amanda R. De La Torre. (2020). "Genomic Basis of White Pine Blister Rust Quantitative Disease Resistance and Its Relationship with Qualitative Resistance." *The Plant Journal* 104, no. 2: 365–76. https://doi.org/10.1111/tpj.14928.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure Author (s): B . S . Weir and C . Clark Cockerham Reviewed work (s): Published by : Society for the Study of Evolution Stable URL : http://www.jstor.org/stable/ 2408641 . *Evolution*, *38*(6), 1358–1370. doi: 128.103.149.52

Williams, B. K., & Brown, E. D. (2019). Sampling and analysis frameworks for inference in ecology. *Methods in Ecology and Evolution*, *10*(11), 1832–1842. doi: 10.1111/2041-210X.13279

Willing, E.-M., Bentzen, P., Oosterhout, C. V., Hoffmann, M., Cable, J., Breden, F., … Dreyer, C. (2010). Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology*, *19*(5), 968–984. doi: 10.1111/j.1365-294X.2010.04528.x

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., … Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, *570*(7762), 514–518. doi: 10.1038/s41586-019-1310-4

Yang, Jinliang, Jiang, H., Yeh, C.-T., Yu, J., Jeddeloh, J. A., Nettleton, D., & Schnable, P. S. (2015). Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. *The Plant Journal*, *84*(3), 587–596. doi: 10.1111/tpj.13029

Yoo, A. B., Jette, M. A., & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. In D. Feitelson, L. Rudolph, & U. Schwiegelshohn (Eds.), *Job Scheduling Strategies for Parallel Processing* (pp. 44–60). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., & Price, A. L. (2013). Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genetics*, *9*(5). doi: 10.1371/journal.pgen.1003520

Zappa Nardelli, F., Belyakova, J., Pelenitsyn, A., Chung, B., Bezanson, J., & Vitek, J. (2018). Julia Subtyping: A Rational Reconstruction. *Proc. ACM Program. Lang.*, *2*(OOPSLA), 113:1–113:27. doi: 10.1145/3276483

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association

studies. *Nature Genetics*, *44*(7), 821–824. doi: 10.1038/ng.2310

Zhu, Y., Bergland, A. O., González, J., & Petrov, D. A. (2012). Empirical Validation of Pooled Whole Genome Population Re-Sequencing in Drosophila melanogaster. *PLoS ONE*, *7*(7), e41901. doi: 10.1371/journal.pone.0041901