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

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

[1]School of Biosciences, [2]Melbournce 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. Here we determine through simulation, the sampling scheme that maximises the power to map the genetic basis of a complex trait 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 performing a shallow sampling of more populations over the landscape which is done best using pool sequencing. Populations should be collected from areas of high genetic diversity and we recommend against sampling from the margins of the species' range. 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

Connecting a species' molecular variation to functional traits is a central goal in ecology. Genomic information for a species can then be leveraged to understand and eventually predict population fitness under a range of eco-evolutionary scenarios. Unfortunately, the required 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 accurately 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 phenotype and genotype data over the whole species' range 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 like restriction site-associated DNA sequencing (RADseq) (Baird et al., 2008) for large 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; Boitard et al., 2012; Cheng et al., 2012; Nielsen et al., 2018), evolution (Boitard et al., 2012; 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. Identifying when best to use one over the other is important. Pool-seq has been shown to accurately capture genome-wide allele frequencies (Fracassetti et al., 2015; Gautier et al., 2013; Rellstab et al., 2013; Zhu et al., 2012), but is also prone to biases in genome representation 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 (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.

Field researchers often transfer techniques initially developed for model organisms or crops in highly controlled environments; however, with natural populations having evolved in natural environments this problem 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 (Indi-seq instead of 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

Landscape simulations

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, and selection intensity with 3 levels for each of these parameters (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 cumulated 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 have q_0=0.01. Under the unidirectional gradient, one boundary row has q_0=0.01 for all 10 populations in that row and 0 elsewhere. Under the bidirectional gradient, two opposite boundary rows have 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

$$1/w = 1 + e^{y_c - y},$$

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 trait values, and *s* is the selection intensity.

The 10,000 biallelic loci were randomly distributed across a genome with 7 chromosomes and a total length of 2×10^9 bases and 750 centimorgans to represent a large genome. The final (200^{th}) generation was sampled for the GPAS. Phenotypic values in this generation were scaled in the 0 to 1 interval for the GPAS experiments.

Genome-wide association and trait prediction based on polygenic scores

GPAS experiments were performed using Indi-seq and Pool-seq data without genotyping error. We used established tools for Indi-seq data, and developed a suite of tools for Pool-seq data. GPAS was performed on all the populations with 384 individuals per population for Indi-seq simulating four 96-well sample plates used in high-throughput molecular biology workflows; and 5 pools per population for Pool-seq, where each pool consists 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.

Allele effects were estimated using 6 Indi-seq-based GPAS (Indi-GPAS) and 3 Pool-seq-based GPAS (Pool-GPAS) modelling frameworks. 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) together with GCTA-derived sparse genetic relationship matrix (GRM; off diagonals <0.05 were set to zero) (Jiang et al., 2019; Yang et al., 2010; Zaitlen et al., 2013) and GEMMA-derived standardised relatedness matrix (STD) (Zhou & Stephens, 2012) were used to build the Indi-GPAS models. 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) were used to build the Pool-GPAS models, with random effect variance determined by F_{ST} derived using Hivert's (Hivert et al., 2018) or Weir and Cockerham's method (Weir & Cockerham, 1984), and variance components estimated using restricted maximum likelihood (REML).

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

152 $$s_{train} = X_{train}\beta,$$

153 where X_{train} is the allele dosage or allele frequency data of the training set, and β are the
154 single-SNP effect estimates. The polygenic scores and actual phenotype values have a linear
155 relationship (Figure S1; mean $R^2_{adjusted}$=0.97±0.0031 at 1,000 individuals per population for Indi-
156 seq and mean $R^2_{adjusted}$=0.99±0.0006 at 5 pools per population for Pool-seq) as expected under
157 the additive model used to simulate the phenotypes. These polygenic scores were regressed
158 against the actual phenotype values of the training set (y_{train}),

159 $$y_{train} = \alpha_0 + \alpha_1 s_{train},$$

160 where α_0 is the intercept, and α_1 is the slope. The polygenic scores of the validation set,
161 $s_{valid} = X_{valid}\beta$, were transformed into the predicted phenotype values ($y_{predicted}$) using

162 $$y_{predicted} = \alpha_0 + \alpha_1 s_{valid}.$$

163

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

166 $$y_{predicted} = X_{valid}\beta,$$

where X_{valid} is the matrix of allele frequencies of the validation set, and $\hat{\beta}$ 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:

$$RMSE = \sqrt{\sum \frac{(y - \hat{y})^2}{n}},$$

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 using all combinations of the 4 landscape parameters allowed to vary 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. To simulate the stratified sampling strategy commonly used in ecology (Hoel, 1943; Li et al., 2017; Williams & Brown, 2019), the landscape was divided into equally sized rectangular regions, and the approximately central population was selected from each region. 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 populations sampled. AUC was used to measure the accuracy of QTL detection per population, while TPR and FPR were used to measure QTL detection accuracy of multiple populations.

The single best performing modelling framework was identified for each genotyping scheme (Indi-GPAS and Pool-GPAS) based on AUC and RMSE for independent populations tests using Tukey's honest significant difference (HSD mean comparison) at $\alpha=0.05$.

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

To determine how many populations to sample to yield optimal GPAS performance, we used the inter-population dataset. We assessed the suitability of the four metrics (i.e. mean AUC, mean RMSE, TPR and FPR) to address this question by visualising their relationships with the number of populations sampled. Additionally, we compared the expected performance of Indi-GPAS and Pool-GPAS under the same sequencing capacity constraint. The Indi-GPAS experiments we simulated included 384 individuals per population, while Pool-GPAS included only 5 pools per population. Assuming a 5X sequencing depth per individual for Indi-seq (Brouard et al., 2017) and the recommended 50X depth per pool for Pool-seq (Schlötterer et al., 2014), these equate

to a sequencing depth of 1,920X per base per population for Indi-seq and only 250X for Pool-seq. This means that for the sequencing capacity required to characterise one population through Indi-seq, approximately 7 populations (⌊1920/250⌋) could be characterised through Pool-seq.

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

The second main 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 used together with 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, we used AUC and RMSE as the GPAS metrics and the intra-population dataset to test the effect of the three causal allele diffusion gradients. The populations were classified into 10 groups, where each group represents a row perpendicular to the causal allele diffusion gradient. The top row refers to populations 1 to 10, the second row to populations 11 to 20, and so on. The general trends and landscape parameter-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 parameters: 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 quantiNemo 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 representatives and the relationships of GPAS performance with landscape and sampling parameters

GEMMA (STD) and GWAlpha showed the best GPAS performances, with >79% AUC and <5.9% RMSE (Table S1). 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 was unaffected by increasing the number of populations sampled since each model was trained independently for each population. In addition, power and accuracy are high when QTL diffuses across the landscape uniformly.

How many populations do we need to sample for optimal GPAS performance? And when should we 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 coincide 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 relationships between QTL detection or prediction accuracies and the sampling location (defined as rows perpendicular to the diffusion gradient) appear 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.

These trends in GPAS performance across the landscape under variable parameter levels correlate with the trends in genetic variability (expressed in terms of causal allele frequency, i.e. frequencies closer to 0.5 indicates higher diversity) across the landscape (Figures S4 to S9). The opposite trends observed between causal allele diffusion gradients for RMSE as selection intensity increases is explained by the shift of the optimal row. At low selection intensity under unidirectional causal allele diffusion, the relationship between RMSE and the row group is linear, i.e. sampling near the diffusion front is better than sampling the middle rows (Figure S5). At high selection intensity under bidirectional causal allele diffusion, the rows in between the middle and top rows, as well as in between the middle and bottom rows become the optima (Figure S8).

Discussion

GPAS in ecology and evolution

GPAS has the potential to expand 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). By focusing on functional traits, GWAS can identify QTL controlling fitness and other ecologically important traits in natural populations. GP exploits the same modelling framework as GWAS to predict phenotype values for the rapid monitoring of species adaptation to existing or changing environmental conditions. The predicted

environmental range of individuals using genome-environment associations (Manel et al., 2018) can be complemented by the phenotype predictions of GPAS. This can be transformational for the way we monitor invasive species or assess the adaptive potential of endangered species.

Our results complement previous research on sampling optimisation in ecology and evolution. We specifically focused on providing recommendations on the optimal sampling strategy to maximise the power to detect QTL and the accuracy of quantitative phenotype prediction in natural populations. 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 here considered phenotyping costs, but higher costs would increase the attractiveness of Indi-seq to maximise information per unit cost. However, similar to a meta analysis of several landscape genomics studies (Santos & Gaiotto, 2020) we demonstrate the value of 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.

Indi-seq provides 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 yield accurate basecalling information (Brouard et al., 2017), and 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 for 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. Hundreds of individuals can be pooled to yield accurate allele frequency data (Schlötterer et al., 2014). This means that a few pools consisting of hundreds of individuals each can represent a population better than a few individual sequences. As expected, 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.

The GPAS models used in this study are representative of quantitative genetics modelling frameworks utilising Indi-seq and Pool-seq genomic data. GEMMA using the standardised relatedness matrix is routinely used for association studies in crops (Begum et al., 2015; Wang et al., 2016; Xiao et al., 2017), livestock (Li et al., 2019; Smith et al., 2019; Wu et al., 2019), and humans (Charng et al., 2020; Fatumo et al., 2019). GWAlpha is a parametric method for estimating additive allelic effects from Pool-seq data that goes beyond the use of just two extreme pools. In this study we have shown that allelic effects estimated using GWAlpha can be used to predict trait values as accurately as Indi-GPAS. In ecological and evolutionary context, Indi-GPAS is performed mostly on model species, e.g. *Arabidopsis thaliana* and *Drosophila melanogaster*, because of the wealth of individual genomic information readily available (1001 Genomes Consortium, 2016; Exposito-Alonso et al., 2018; Flatt, 2020; Mackay et al., 2012). The

cost-effectiveness of Pool-seq can help close the gap in the application of these powerful statistical frameworks between model and non-model species. Hence, Pool-GPAS with GWAlpha can bring a powerful and cost-effective framework to ecology and evolution for the identification of the genetic basis of quantitative traits and the prediction of trait values for the rapid monitoring of species adaptation.

GPAS performance as affected by trait polygenicity and landscape properties

QTL detection and phenotype prediction accuracies can have similar or contrasting responses to varying genetic architectures and landscape properties. QTL detection power increases as the contribution of each QTL to the trait increases and as the frequency of QTL alleles is balanced (ie. close to 0.5). On the other hand, prediction accuracy increases as the total additive genetic variance increases. Allele frequencies and additive genetic variance vary in response to evolutionary forces, e.g. selection intensity and migration rate. Some prior knowledge of the polygenicity of the trait and the spatial heterogeneity of selection intensity and gene flow is valuable for identifying the optimal locations for population sampling.

There is less power to detect QTL but higher prediction accuracy in highly polygenic traits than traits controlled by fewer loci. Increasing the number of loci controlling the trait reduces the selection pressure acting on each locus (Walsh & Lynch, 2018). This decreases the power to detect QTL since the individual contribution of each QTL decreases as more loci control the trait (Wang & Xu, 2019). This in turn reduces the proportion of polymorphic QTL within populations: if the majority of the QTL have small effects, they have a higher chance of getting lost due to drift than QTL with large effects. On the other hand, prediction accuracy increases since the rate at

which genetic variance decreases due to directional selection is reduced as the number of QTL increases. Genetic variance should eventually become zero under constant stabilising or directional selection, but the rate of this reduction becomes slower with an increased number of loci controlling the trait (Crow & Kimura, 1970). Hence, GWAS and GP complement each other to achieve high QTL detection accuracy or high prediction accuracy for quantitative traits controlled by any number of loci.

There is higher power to detect QTL but lower prediction accuracy in populations under intense selection. Increasing the selection intensity increases QTL detection power, since the contribution of individual QTL becomes greater in each population (Wang & Xu, 2019) and less QTL alleles are lost due to drift. However, the predictive ability will be reduced by the Bulmer effect (Bulmer, 1971), where covariance between loci (partially explained by linkage disequilibrium (Walsh & Lynch, 2018)) is reduced after selection. Increasing 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.

There is more power to detect QTL and greater prediction accuracy in populations experiencing high migration than in reproductively isolated ones. Increasing migration rate decreases differentiation between populations allowing for more causal alleles to be shared, resulting in higher power and more accurate predictions. This is expected to reduce the contribution of each QTL since the number of QTL per population increases thereby decreasing power (Griswold, 2006; Wang & Xu, 2019). However, our results suggest the opposite: power increases as

migration increases. This is because the total number of loci controlling the trait does not increase per se, only the proportion of the polymorphic QTL per population does, which leads to higher variance for these QTL per population. Increasing the additive genetic variance in each population (Liu et al., 2020) also results in higher prediction accuracy.

Sampling strategy for GPAS in ecology and evolution

The results of this simulation study 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 highly 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 sampling an increasing number of populations. This is consistent with a previous study on sampling strategy optimisation which found that sampling an intermediate number of sites can perform as well as maximising the number of sites sampled (Selmoni et al., 2020). 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.

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 Pool-seq and more populations can be characterised with Pool-seq than 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.

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. In the context of a species distributed over a restricted range, alleles have a higher probability of getting lost in border populations than in the populations in the middle of the range. The non-linear relationship between the RMSE and the distance from the diffusion front under unidirectional diffusion gradient also highlights this border effect. Populations in the middle of the distribution range have a similar number of polymorphic causal loci as the populations closer to the diffusion front. This phenomenon reflects what happens 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). Hence, in the absence of prior information on the location of high genetic diversity, we do not recommend sampling from these fringe or border populations.

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

Acknowledgments

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.

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

1001 Genomes Consortium, J., 1001 Genomes Consortium. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, *166*(2), 481–491. doi: 10.1016/j.cell.2016.05.063

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

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

Begum, H., Spindel, J. E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., … McCouch, S. R. (2015). Genome-Wide Association Mapping for Yield and Other Agronomic Traits in an Elite Breeding Population of Tropical Rice (Oryza sativa). *PLoS ONE*, *10*(3). doi: 10.1371/journal.pone.0119873

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

Boitard, S., Schlotterer, C., Nolte, V., Pandey, R. V., & Futschik, A. (2012). Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. *Molecular Biology and Evolution*, *29*(9), 2177–2186. doi: 10.1093/molbev/mss090

Brouard, J.-S., Boyle, B., Ibeagha-Awemu, E. M., & Bissonnette, N. (2017). Low-depth

584 genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the
585 selection of high quality genotypes and the accuracy of imputation. *BMC Genetics*, *18*.
586 doi: 10.1186/s12863-017-0501-y
587
588 Bulmer, M. G. (1971). The Effect of Selection on Genetic Variability. *The American Naturalist*.
589 Retrieved from https://www.journals.uchicago.edu/doi/10.1086/282718
590
591 Cappa, E. P., El-Kassaby, Y. A., Garcia, M. N., Acuña, C., Borralho, N. M. G., Grattapaglia, D., &
592 Poltri, S. N. M. (2013). Impacts of Population Structure and Analytical Models in
593 Genome-Wide Association Studies of Complex Traits in Forest Trees: A Case Study in
594 Eucalyptus globulus. *PLOS ONE*, *8*(11), e81267. doi: 10.1371/journal.pone.0081267
595
596 Charng, J., Simcoe, M., Sanfilippo, P. G., Allingham, R. R., Hewitt, A. W., Hammond, C. J., …
597 Yazar, S. (2020). Age-dependent regional retinal nerve fibre changes in SIX1/SIX6
598 polymorphism. *Scientific Reports*, *10*. doi: 10.1038/s41598-020-69524-8
599
600 Cheng, C., White, B. J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M. W., & Besansky, N.
601 J. (2012). Ecological genomics of anopheles gambiae along a latitudinal cline: A
602 population-resequencing approach. *Genetics*, *190*(4), 1417–1432. doi:
603 10.1534/genetics.111.137794
604
605 Crow, J. F., & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. doi:
606 10.2307/1529706
607
608 Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, *186*(1), 41–3. doi:
609 10.1534/genetics.110.121012
610
611 de Roos, A. P. W., Hayes, B. J., & Goddard, M. E. (2009). Reliability of genomic predictions
612 across multiple populations. *Genetics*, *183*(4), 1545–53. doi:
613 10.1534/genetics.109.104935
614
615 Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., … Singer, A.
616 (2012). Correlation and process in species distribution models: bridging a dichotomy.
617 *Journal of Biogeography*, *39*(12), 2119–2131. doi: 10.1111/j.1365-2699.2011.02659.x
618
619 Eckert, A. J., & Dyer, R. J. (2012). Defining the landscape of adaptive genetic diversity.
620 Molecular Ecology, 21(12), 2836–2838. doi: 10.1111/j.1365-294X.2012.05615.x
621
622 Edwards, S. M., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., … Hickey, J. M.
623 (2019). The effects of training population design on genomic prediction accuracy in
624 wheat. *Theoretical and Applied Genetics*, *132*(7), 1943–1952. doi: 10.1007/s00122-019-
625 03327-y
626
627 Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H. A., & Weigel, D. (2018).
628 Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis
629 thaliana. *Nature Ecology & Evolution*, *2*(2), 352–358. doi: 10.1038/s41559-017-0423-0

Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C., … SanCristobal, M. (2017). Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: The local score approach. *Molecular Ecology*, *26*(14), 3700–3714. doi: 10.1111/mec.14141

Fatumo, S., Carstensen, T., Nashiru, O., Gurdasani, D., Sandhu, M., & Kaleebu, P. (2019). Complimentary Methods for Multivariate Genome-Wide Association Study Identify New Susceptibility Genes for Blood Cell Traits. *Frontiers in Genetics*, *10*. doi: 10.3389/fgene.2019.00334

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi: 10.1016/j.patrec.2005.10.010

Flatt, T. (2020). Life-History Evolution and the Genetics of Fitness Components in Drosophila melanogaster. *Genetics*, *214*(1), 3–48. doi: 10.1534/genetics.119.300160

Fournier-Level, A., Korte, A., Cooper, M. D., Nordborg, M., Schmitt, J., & Wilczek, A. M. (2011). A Map of Local Adaptation in Arabidopsis thaliana. *Science*, *334*(6052), 86–89. doi: 10.1126/science.1209271

Fournier-Level, Alexandre, Robin, C., & Balding, D. J. (2017). GWAlpha: Genome-wide 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/

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

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

Griswold, C. K. (2006). Gene flow's effect on the genetic architecture of a local adaptation and its consequences for QTL analyses. *Heredity*, *96*(6), 445–453. doi:

676 10.1038/sj.hdy.6800822
677
678 Guo, X., Cericola, F., Fè, D., Pedersen, M. G., Lenk, I., Jensen, C. S., … Janss, L. L. (2018).
679 Genomic Prediction in Tetraploid Ryegrass Using Allele Frequencies Based on
680 Genotyping by Sequencing. *Frontiers in Plant Science*, *9*, 1165. doi:
681 10.3389/fpls.2018.01165
682
683 Hansson, B., Sigeman, H., Stervander, M., Tarka, M., Ponnikas, S., Strandh, M., … Hasselquist,
684 D. (2018). Contrasting results from GWAS and QTL mapping on wing length in great
685 reed warblers. *Molecular Ecology Resources*, *18*(4), 867–876. doi: 10.1111/1755-
686 0998.12785
687
688 Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring Genetic
689 Differentiation from Pool-seq Data. *Genetics*, *210*(1), 315–330. doi:
690 10.1534/genetics.118.300900
691
692 Hoel, P. (1943). The Accuracy of Sampling Methods in Ecology on JSTOR. *The Annals of*
693 *Mathematical Statistics*, *14*(3), 289–300.
694
695 Holliday, J. A., Ritland, K., & Aitken, S. N. (2010). Widespread, ecologically relevant genetic
696 markers developed from association mapping of climate-related traits in Sitka spruce
697 (Picea sitchensis). *New Phytologist*, *188*(2), 501–514. doi: 10.1111/j.1469-
698 8137.2010.03380.x
699
700 Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A
701 resource-efficient tool for mixed model association analysis of large-scale data. *Nature*
702 *Genetics*, *51*(12), 1749–1755. doi: 10.1038/s41588-019-0530-8
703
704 Johnston, S. E., McEWAN, J. C., Pickering, N. K., Kijas, J. W., Beraldi, D., Pilkington, J. G., …
705 Slate, J. (2011). Genome-wide association mapping identifies the genetic basis of
706 discrete and quantitative variation in sexual weaponry in a wild sheep population.
707 *Molecular Ecology*, *20*(12), 2555–2566. doi: 10.1111/j.1365-294X.2011.05076.x
708
709 Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., … Eskin, E.
710 (2010). Variance component model to account for sample structure in genome-wide
711 association studies. *Nature Genetics*, *42*(4), 348–354. doi: 10.1038/ng.548
712
713 Knight, J., Saccone, S. F., Zhang, Z., Ballinger, D. G., & Rice, J. P. (2009). A Comparison of
714 Association Statistics between Pooled and Individual Genotypes. *Human Heredity*, *67*(4),
715 219–225. doi: 10.1159/000194975
716
717 Lafayette, L., & Wiebelt, B. (2017). Spartan and NEMO: Two HPC-Cloud Hybrid
718 Implementations. *2017 IEEE 13th International Conference on E-Science (e-Science)*,
719 458–459. Auckland: IEEE. doi: 10.1109/eScience.2017.70
720
721 Li, X., Nie, C., Liu, Y., Chen, Y., Lv, X., Wang, L., … Qu, L. (2019). A genome-wide association

study explores the genetic determinism of host resistance to Salmonella pullorum infection in chickens. *Genetics, Selection, Evolution : GSE*, *51*. doi: 10.1186/s12711-019-0492-4

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

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

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., … Gibbs, R. A. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature*, *482*(7384), 173–178. doi: 10.1038/nature10811

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

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

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

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

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

Smith, J. L., Wilson, M. L., Nilson, S. M., Rowan, T. N., Oldeschulte, D. L., Schnabel, R. D., … Seabury, C. M. (2019). Genome-wide association and genotype by environment interactions for growth traits in U.S. Gelbvieh cattle. *BMC Genomics*, *20*(1), 926. doi:

10.1186/s12864-019-6231-y

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

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, H., Xu, X., Vieira, F. G., Xiao, Y., Li, Z., Wang, J., … Chu, C. (2016). The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication. *Molecular Plant*, *9*(7), 975–985. doi: 10.1016/j.molp.2016.04.018

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

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

Wu, P., Wang, K., Zhou, J., Yang, Q., Yang, X., Jiang, A., … Tang, G. (2019). A genome wide association study for the number of animals born dead in domestic pigs. *BMC Genetics*, *20*(1), 4. doi: 10.1186/s12863-018-0692-x

Xiao, Y., Liu, H., Wu, L., Warburton, M., & Yan, J. (2017). Genome-wide Association Studies in Maize: Praise and Stargaze. *Molecular Plant*, *10*(3), 359–374. doi: 10.1016/j.molp.2016.12.008

Yang, Jian, Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. doi: 10.1038/ng.608

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

884 # Tables and Figures
885
886 **Table 1**. List of parameters used in landscape simulations.

Parameter	Levels
Number of populations per landscape	100
Number of hermaphroditic individuals per population	1,000
Number of loci per individual	10,000
Number of generations	200
Number of loci controlling the trait	10, 50, and 100
Migration rate per population per generation	0.0001, 0.001, and 0.01
Causal allele diffusion gradient	Uniform, unidirectional, and bidirectional
Selection intensity	0.50, 0.90, and 0.95
Number of replicates per landscape	5

887

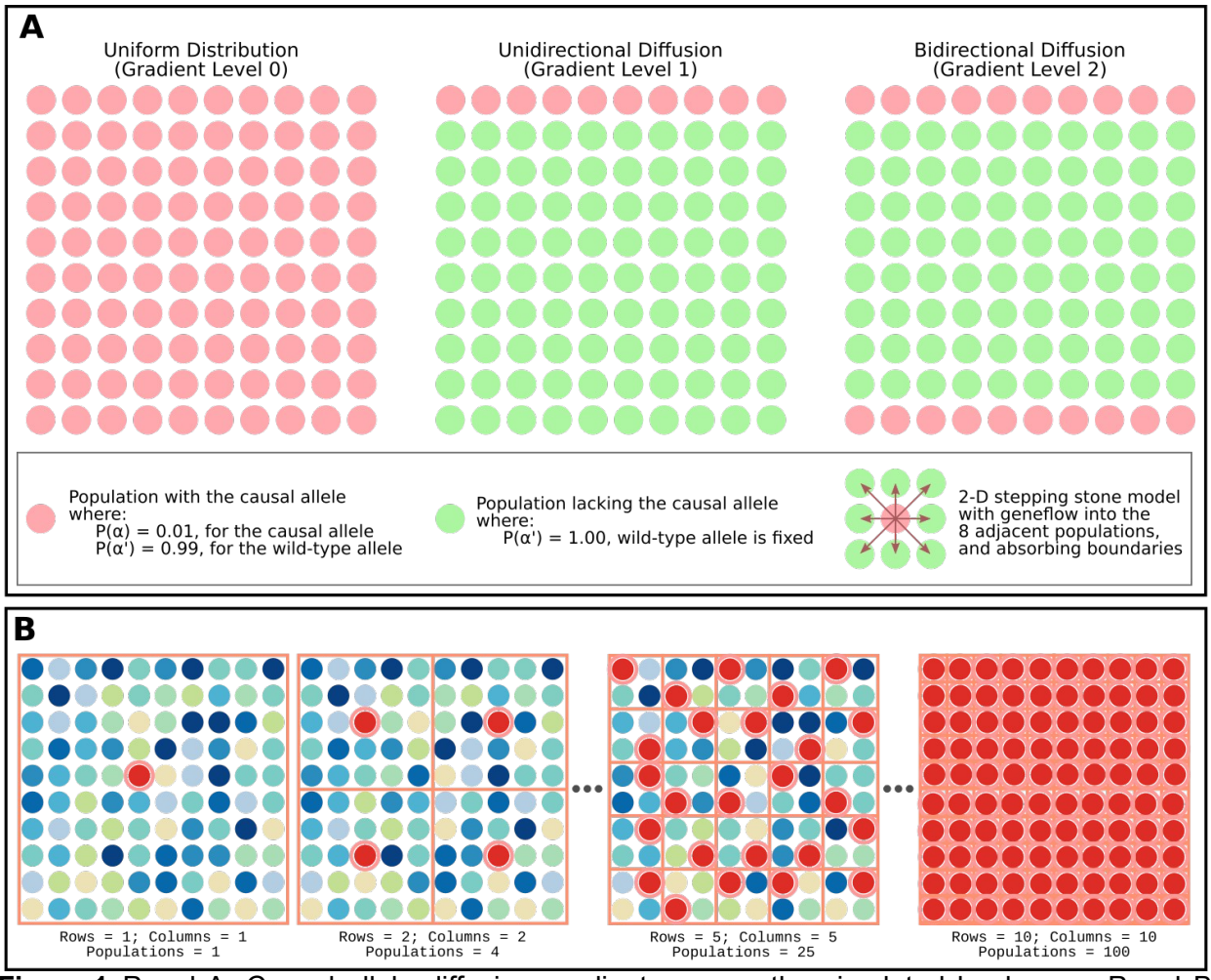

Figure 1. Panel A: Causal allele diffusion gradients across the simulated landscape. Panel B: Systematic sampling strategy across the simulated landscape. Note: For one population sampling, all populations across the landscape were sampled independently.

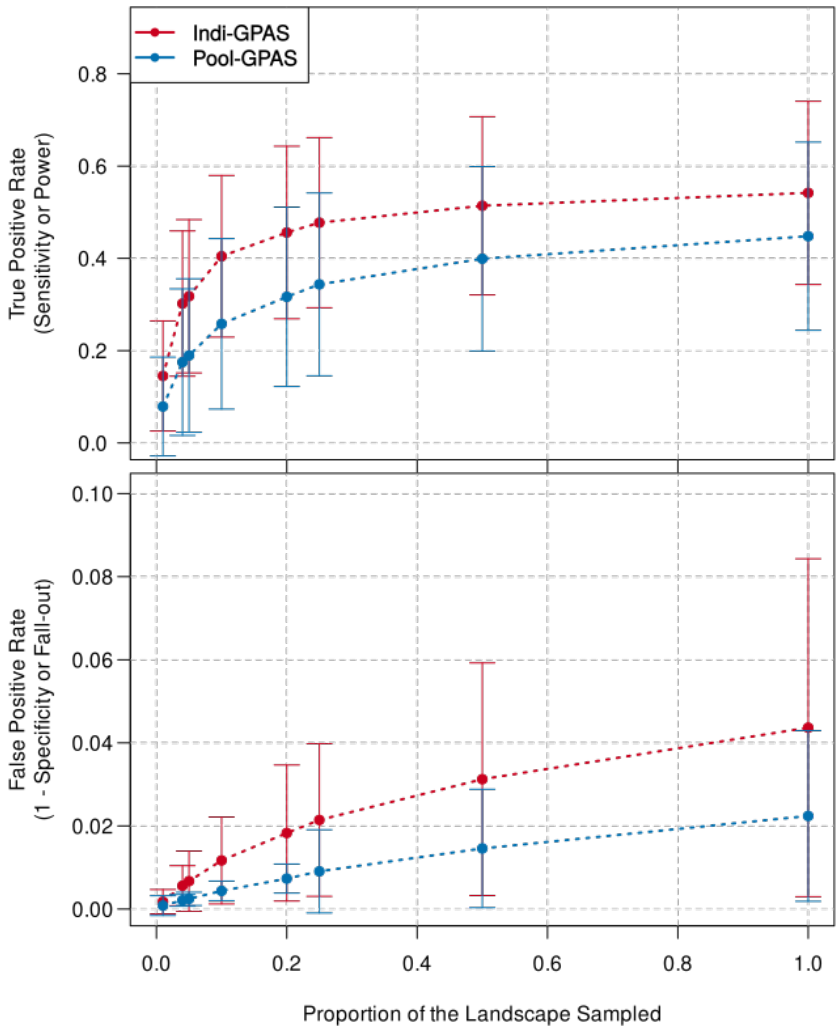

Figure 2. Relationships between the proportion of the landscape sampled and GPAS performance. Dots represent means and whiskers indicate ±1 standard deviation from the mean.

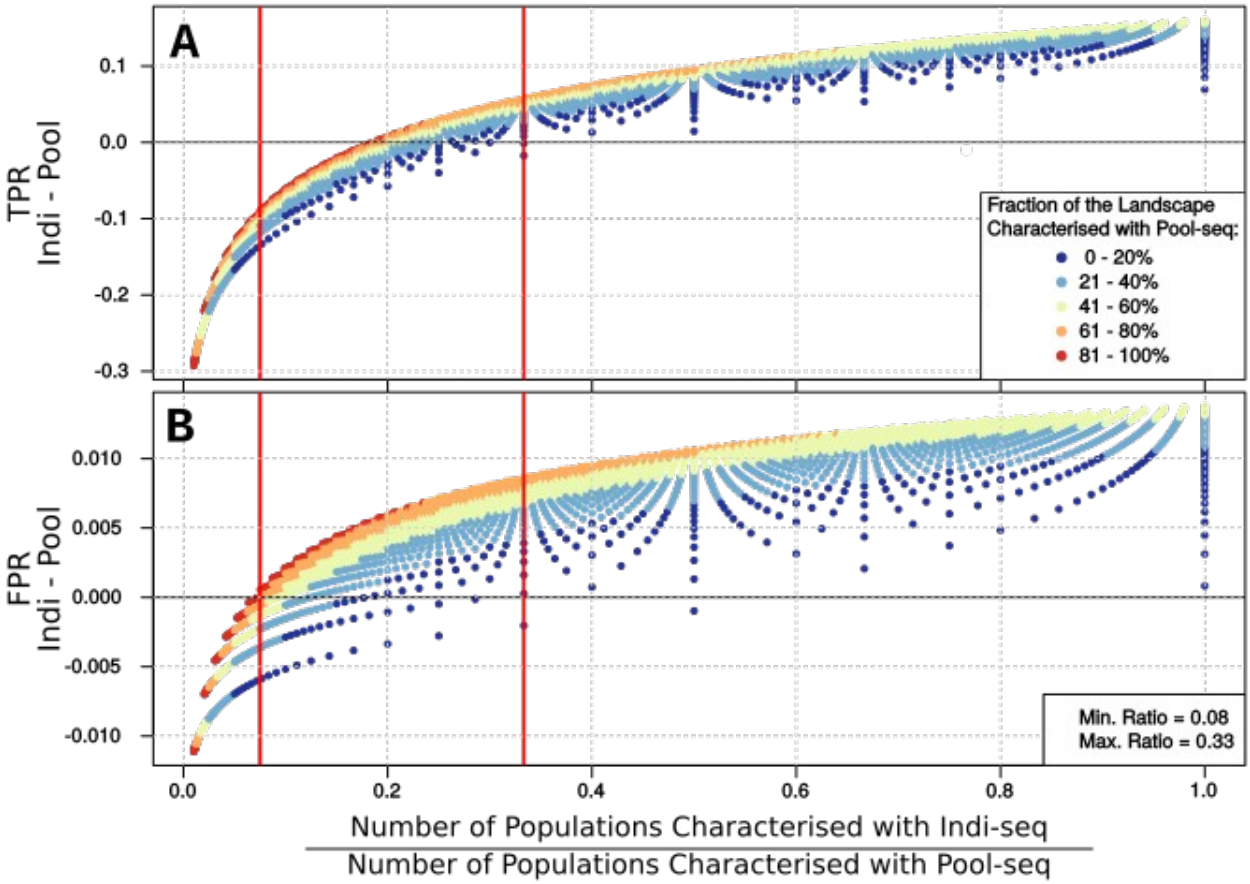

Figure 3. Pool-GPAS can outperform Indi-GPAS under the same sequencing capacity. **Panel A**: Difference in true positive rate (TPR) between Indi-GPAS and Pool-GPAS models as the ratio between the number of populations characterised through Indi-seq and Pool-seq increases. **Panel B**: Difference in false positive rate (FPR) between Indi-GPAS and Pool-GPAS models as the ratio between the number of populations characterised through Indi-seq and Pool-seq increases. **Vertical red lines**: The range of ratios between the two vertical red lines correspond to cases when Pool-GPAS can outperform Indi-GPAS in our study, i.e. Pool-GPAS has higher TPR and lower FPR than Indi-GPAS. Moving along the x-axis involves modifying the sequencing depths for Indi-seq or Pool-seq and not the number of individuals or pools sampled per population, since 384 individuals per population and 5 pools per population were kept constant in the simulations.

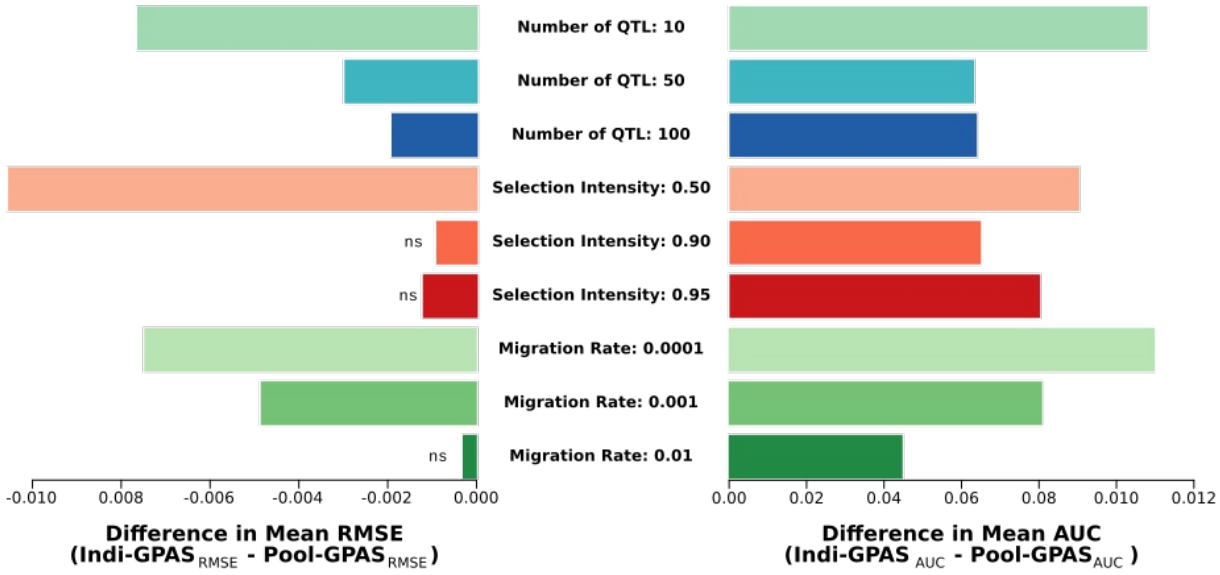

Figure 4. Differences in QTL detection and polygenic score prediction accuracies between Indi-GPAS and Pool-GPAS per population. AUC = Area under the ROC curve. RMSE = root mean square error of the polygenic score prediction. The term "*ns"* beside the bars indicates non-significant differences based on HSD at p<0.05.

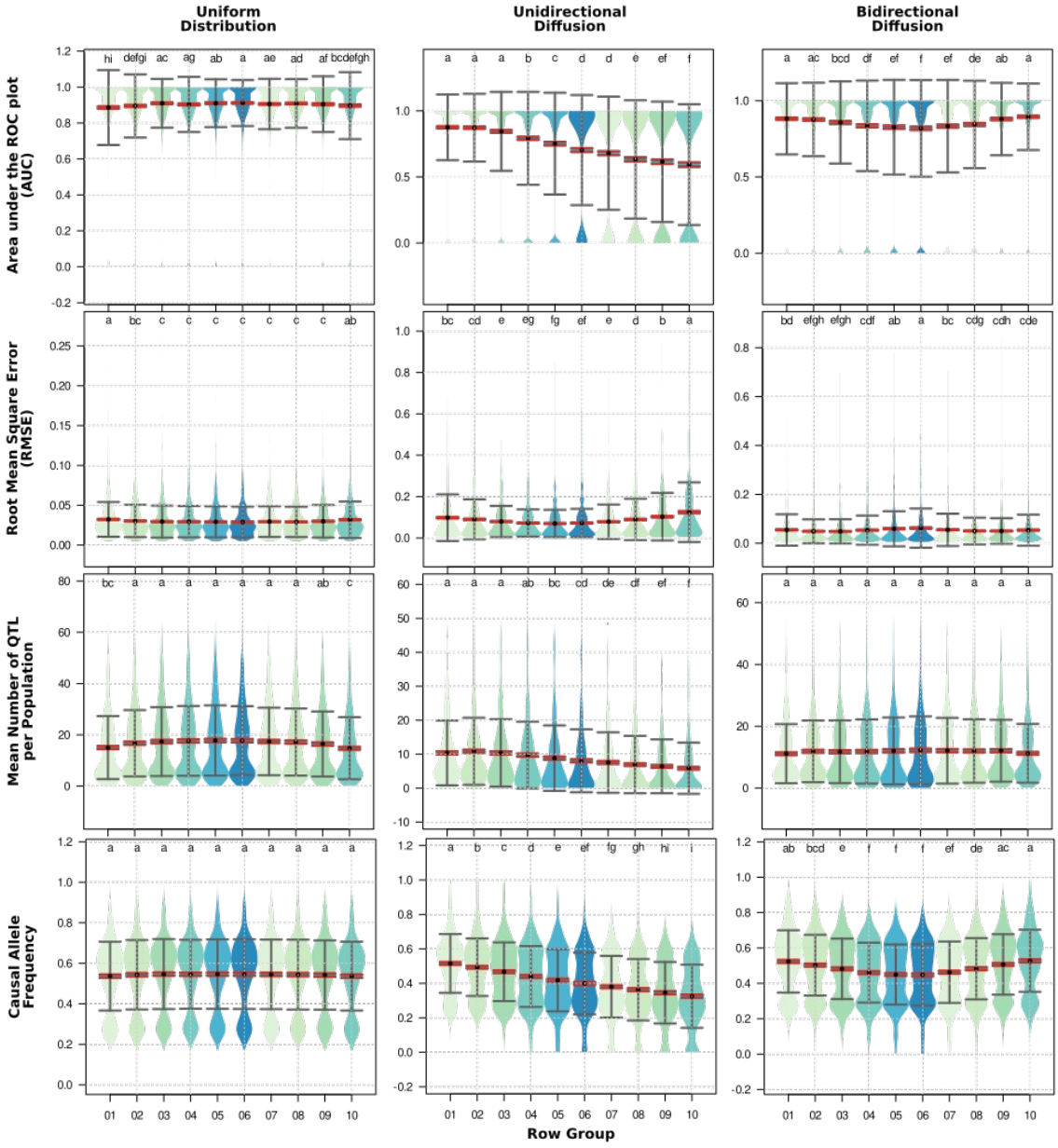

Figure 5. Violin plots of the area under the receiver operating curve (AUC; measure of QTL detection accuracy), root mean square error (RMSE; measure of polygenic score prediction accuracy), mean number of polymorphic causal loci per population, and causal allele frequencies across the landscape divided into 10 rows under uniform, unidirectional, and bidirectional causal allele diffusion. Each row is perpendicular to the causal allele diffusion gradient. Black dots represent the mean, black whiskers show ±1 standard deviation, red whiskers are the 95% confidence interval, and the letters on top of each plot are Tukey's honest significant difference (HSD)-based grouping (i.e. row groups with the same letter are not significantly different at $p<0.05$).

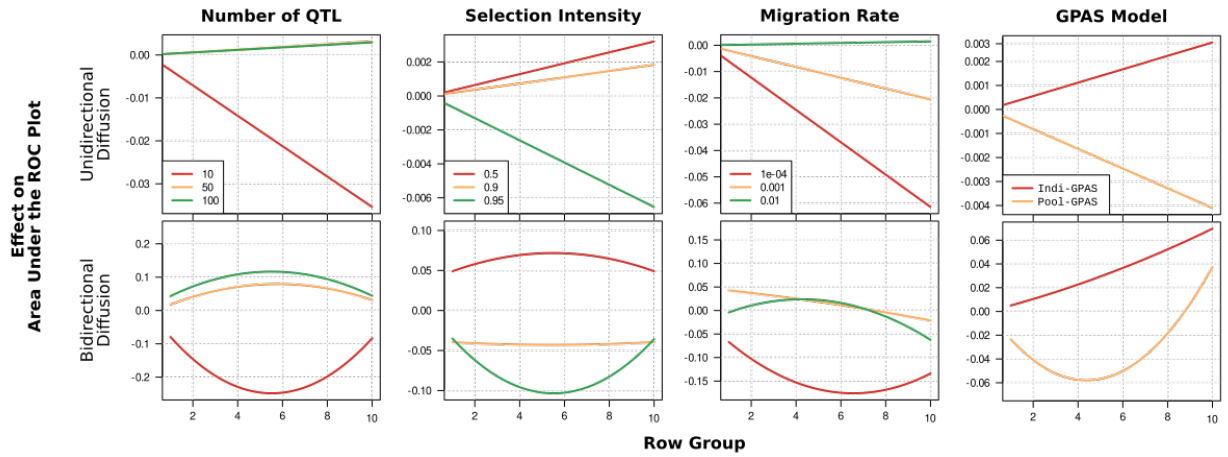

Figure 6. Effects of sampling across the landscape on QTL detection accuracy under varying number of simulated QTL, selection intensities, migration rates, and GPAS models for unidirectional (top graphs) and bidirectional (bottom graphs) causal allele diffusion gradients. QTL detection accuracy was measured using the area under the curve which describes the relationship between power and false positive rate, where high values mean high QTL detection accuracies.

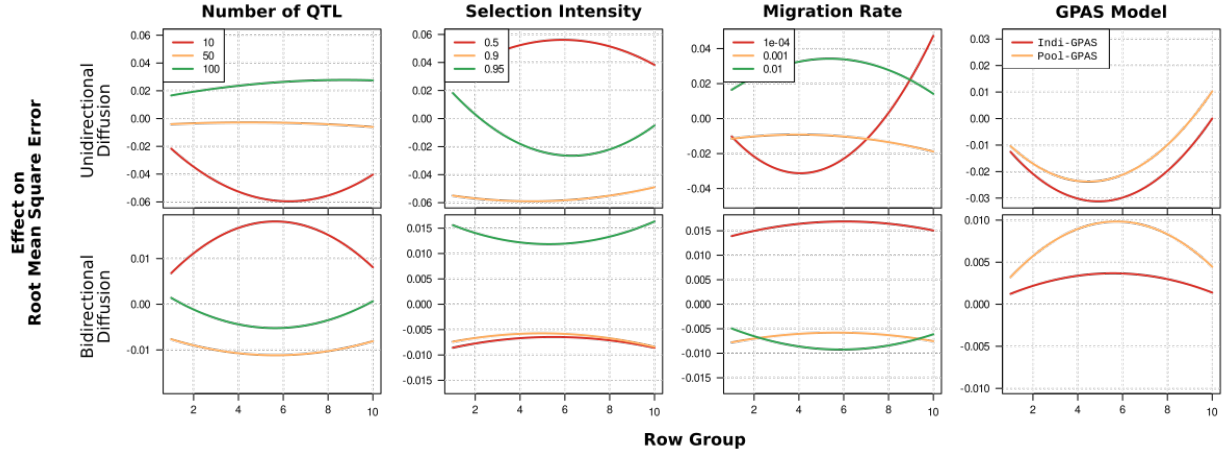

Figure 7. Effects of sampling across the landscape on prediction accuracy under varying number of simulated QTL, selection intensities, migration rates, and GPAS models for unidirectional (top graphs) and bidirectional (bottom graphs) causal allele diffusion gradients. Polygenic score prediction accuracy was measured using root mean square error where low values mean high accuracies.